This 2025 work is licensed under CC BY-NC-SA 4.0

Quantitative Sociology Essentials: An Introduction to Survey Data Analysis using IBM® SPSS®

Ed Collom, Ph.D.
Professor of Sociology
California State University, Fullerton

Table of Contents

Chapter	*Title*	*Page*
1	The Fundamentals of Sociological Survey Research	1
2	Inspecting Data with SPSS	11
3	Measures of Central Tendency	35
4	Measures of Dispersion	49
5	Making Inferences	63
6	Estimation	74
7	Hypothesis Testing	85
8	Crosstabulations and Measures of Association	90
9	*t* Testing	110
10	Analysis of Variance	121
11	Regression and Correlation	134

This 2025 work is licensed under CC BY-NC-SA 4.0

Quantitative Sociology Essentials: An Introduction to Survey Data Analysis using IBM® SPSS®

Ed Collom, Ph.D.
Professor of Sociology
California State University, Fullerton

Chapter 1. The Fundamentals of Sociological Survey Research

As many scientists do, quantitative sociologists use procedures to produce numbers that summarize trends in data that have been collected. **Statistics** on empirical observations are everywhere in contemporary society and often serve as the basis for social policy recommendations. Given the ubiquity of information and statistics, as well as their use in political agendas, critical thinking skills are as important as ever. Innumeracy, the inability to work with and understand numbers, limits people's capacity to be engaged and informed citizens.[1] Data analysis and reporting skills are also in high demand in today's labor market.

The goal of this text is to teach undergraduate students to perform basic quantitative sociology. The focus is upon analyzing high-quality, survey data using IBM® SPSS®, a powerful and popular data analysis software package. Students will first learn about descriptive statistics to summarize patterns in survey data. The second half of the text concerns the major inferential statistical techniques that enable sociologists to make generalizations about social phenomena observed in sample surveys to the larger populations that are being represented.

Chapter Objectives

After reading this chapter, students should be able to:
- Summarize the steps of the research process
- Differentiate independent and dependent variables
- Select a variable's level of measurement

The Steps of the Research Process

Quantitative sociologists employ a version of the scientific method to ensure that their research methods are rigorous, objective, and replicable. These steps of the sociological research process are as follows:

- 1) Posing a research question
- 2) Reviewing the literature
- 3) Formulating the research hypothesis
- 4) Collecting the data
- 5) Analyzing the data
- 6) Reporting the findings
- 7) Interpreting the findings
- 8) Sharing the findings

We begin the process by posing our research question. The **research question** is the topic of the study and is often stated as a question about a potential relationship between attributes of the unit of analysis. The **unit of analysis** is the specific entity we want to study. In sociological research involving surveys, the unit of analysis is typically individuals. That is, individual respondents answer survey questions pertaining to their personal characteristics and opinions. A **variable** is any aspect or attribute of a unit of analysis that varies among cases. In sociology, common survey variables include age, sex, race, educational attainment, and income. Thus, a sociological research question frames the topic by asking whether two variables are associated. For example, do those with greater education know more people? Or, alternatively, does the size of one's social networks vary by educational attainment?

The second step of the research process is to review the scholarly literature. A literature review for a quantitative sociology study typically involves sociological theory and previous empirical research. Sociological theory involves assumptions of how societies operate and provides explanations of social phenomena. Theory can provide a strong basis for your expected findings.

For our example, we could consult sociological theory on social capital and social networks. One of the leading theorists in this area argues that those with greater educational attainment are more likely to have access to social ties and therefore, better social capital overall.[2] Thus, sociological theory suggests that those with more education are likely to know more people.

As part of the literature review, it is also crucial to learn what previous researchers may have found on your topic. You will likely not be the first person to have conducted research on this issue. Previous research findings can also be a strong basis for your own anticipated results. One of the most frequently cited journal articles in the discipline of sociology was published in the field's top journal (the *American Sociological Review*) in 2006. Studying two different waves of the same survey that is the basis of this text, the authors find that the number of people with whom one has discussed important matters with over the past six months increases with greater educational attainment.[3] Thus, previous research also indicates that those with more education have been found to have larger social networks.

The intended outcome of the review of theoretical and previous empirical research literatures is the formulation of a research hypothesis (the third step of the process). Research hypotheses are predictions of how one variable will impact another: those with greater educational attainment are expected to have larger social networks. This statement anticipates that the variable educational attainment will influence the variable social network size.

In sociology and in survey research, we often think about social phenomena in terms of cause and effect. **Independent variables** are those that produce an impact upon another variable (or "predictors"). **Dependent variables** are the outcomes or affected variables. Since the dependent variable is the property that you are trying to explain, it is the main focus and topic of the research. As we shall see later, independent variables are also referred to as "X" and dependent variables as "Y."

Ideally, the literature review will result in the identification of the independent and dependent variables. Yet when studying social issues, it is not always clear which variable influences the other. Sociologists must consider whether the proposed relationship is logical. Are there reasons why the independent variable may impact the dependent variable? If you swap the order and the reverse does not make sense, you are likely on the right track. Timing is also important to consider since independent variables ("the cause") must precede dependent variables ("the effect").

Sociologists also think about variables in terms of ascribed versus achieved characteristics.[4] **Ascribed characteristics** are statuses that we inherit or have little or no control over. None of us chose the time and

place at which we were born. Yet sociologists understand that the eras and regions in which we live have particular social structures that can benefit or constrain our potential (or "life chances"[5]). Thus, variables such as age, sex, and race are ascribed statuses that are only treated as independent variables. **Achieved characteristics** are those that we earn or develop. These are statuses over which we are able to exercise some agency and control. While achieved statuses such as educational attainment and income are often treated as outcomes or dependent variables, they also frequently serve as useful predictors.

Let's return to our sample research question: Do those with greater education know more people? Educational attainment would be our independent variable here and social network size our dependent variable. An arrow is often used to symbolize the predicted causation: Educational Attainment → Social Network Size.

A **research hypothesis** is a specific statement of how the independent variable is expected to impact the dependent variable: those with greater educational attainment are expected to have larger social networks. Notice that the direction of the predicted association is implied here as well. That direction is positive: higher educational attainment leads to more social network ties.

After the research hypothesis has been formulated, it is time to gather data that is capable of testing the hypothesis. As may be evident by now, the steps of the sociological research process are not strictly linear. Scholars are typically considering many of them simultaneously as they assess the feasibility of their ideas.

The focus of this text is secondary survey data analysis. **Secondary data** is already in existence and has been collected by someone else. Survey data collection is a science in and of itself and often best left to the professionals. The most robust datasets are those that are based on nationally representative samples of respondents that are randomly selected (samples and sampling methods will be covered later in Chapter 5). **Primary data** is original information that researchers collect for their specific project. While it can definitely be advantageous to write your own survey questions for your topic, high-quality surveys are usually expensive to administer. Surveys fielded by individual researchers tend to be smaller and often have limited response rates.

The main dataset to be employed in this text is the 2021 General Social Survey (GSS) of the National Opinion Research Center (NORC). The GSS is a nationally representative survey of American adults that has been conducted since 1972. It is the most widely used and best source of survey data to investigate sociological and attitudinal trends. Over 4,000 respondents completed the web-based survey in 2021 and there are hundreds of variables available to analyze. The codebook contains the exacting wording of all of the questions as well as all of the methodological details. Nearly all of the questions are close-ended as respondents are presented with response categories from which to choose.

The fourth step of the research process (collecting the data) is easy in secondary survey data research as we need only to access the data files. Over the years, the GSS has included questions about social network ties and has always asked respondents for their educational attainment. Therefore, the GSS would be capable of answering the research question posed earlier and testing the stated research hypothesis.

Analyzing the data, the fifth step of the research process, comprises the bulk and primary goal of this text. As we will see throughout, we must carefully select the appropriate statistical tools based upon the characteristics of the variables that we will be analyzing. We will begin with **descriptive statistics** that summarize survey responses to individual variables (also known as "univariate" statistics). In the second half of the text we will learn the major techniques of testing for associations between independent and dependent variables. These statistical tests are known as **inferential statistics** since they indicate whether or not we can be confident in generalizing any associations between variables ("bivariate" statistics) from the samples that we analyze to the larger populations that are being represented in the data.

The sixth step of the sociological research process is reporting the findings. Here, researchers are required to clearly summarize the results of the statistics that they produce. We will begin by learning to report descriptive statistics such as the mean, the most well-known measure of central tendency. As we progress into inferential statistics, we will report on associations between two variables that are generalizable. After reporting the results, researchers must determine whether their research hypothesis is supported by their findings. The assessment of the hypothesis is key to this step of the research process.

The next step of the research process is interpreting the findings. This is the stage in which sociologists attempt to explain why their results actually exist in society. Why is it that the independent variable produced an impact upon the dependent variable? Returning to our earlier example, let's say that we did find a positive association between educational attainment and social network size. It is now our responsibility to try to explain why. Our literature review comes back into play here to guide our interpretation. If our research hypothesis that was grounded in the literature is supported, we can reiterate that literature in our explanation. As we shall see later, we will also be able to discuss the strength of the association in our interpretation. In some cases, the associations we find may not be as strong as we expected. In other cases, we may find that there is no association between the two variables at all. This situation is the most challenging one for interpretations. We must still interpret our findings and attempt to explain why they are not associated despite our expectations. Non-associations are still findings and may be as important as associations.

The final step of the sociological research process is sharing the findings. Quantitative sociologists write reports, conference presentations, journal articles, and books. We disseminate our findings with our peers in our subfields. If our research is accepted for publication as an article in a scholarly, peer-reviewed academic journal, it is legitimated as an original, methodologically-sound work that makes a contribution. Once scholarship is published, it then becomes part of the existing literature for future researchers to engage with in their projects. Thus, the research process is a cycle and our published work can influence future scholars as they pose their own research questions.

The Levels of Measurement of Variables

Quantitative sociologists employing secondary data typically spend a great deal of time reading survey instruments (codebooks) to locate variables capable of testing their research hypotheses. Statistical analyses are based on the mathematical properties of variables. Our analysis of the survey data is ultimately based upon how those who fielded the survey instrument asked the questions in which we are interested.

As we will see in the next chapter, SPSS only calculates statistics and processes trends among sets of numbers. Any social phenomenon that we want to study quantitatively must be reduced to numbers. Every

variable has multiple **response categories** that reflect how the individuals who completed the survey answered the questions.

Nearly all surveys (and many of the forms that we regularly fill out) ask respondents to report their sex or gender. The GSS collects sex and gender identity in two separate items: sex recorded at birth and current gender identity. The sex at birth survey question is "Was your sex recorded as male or female at birth?" The response categories "male" and "female" are assigned numerical **values** that permit SPSS to process this sex variable. In the GSS datafile that we will analyze, males are coded as "1" and females are coded as "2."

The current gender identity question is "Do you describe yourself as male, female, or transgender?" This variable is coded with the following values: males = 1, females = 2, and transgender = 3. SPSS calculates statistics from the basis of these values, not the text comprising the response categories. The software does not analyze the substance of the variables. That is the sociologist's job.

Every variable has a **level of measurement** that determines the statistical analyses that we can perform on it. **Nominal** variables are the lowest level of the three as they are the most mathematically limited. Response categories in nominal variables are text-based. That is, survey respondents are asked a close-ended question and choose a response category comprised of words that best represents them. In addition to the response categories being words, another defining feature of nominal variables is that those response categories cannot be ranked in any meaningful way. There is no inherent order to the response categories of nominal variables. The sex and the gender variables described above are both examples of nominal variables. The response categories are textual, describing different characteristics, and the categories cannot be ranked. Other examples of nominal variables include race/ethnicity variables as well as religious affiliation.

Ordinal variables also have response categories that are text-based. However, response categories in ordinal variables can be ranked and do have some inherent order to them. Different response categories can be thought of as being higher or lower than others. For example, many surveys provide statements about various issues and ask respondents the extent to which they agree with the statements: "strongly agree," "somewhat agree," "neither agree nor disagree," "somewhat disagree," or

"strongly disagree." This approach to measuring attitudes and opinions through ranked categories is known as a Likert scale. Survey researchers also commonly use the phrase "To what extent…" and then provide a statement for consideration by the respondents. The response categories for this type of question are often: "not at all," "to some extent," or "to a great extent."

In addition to being the most common level of measurement for attitudinal questions, variables that measure the frequency of some activity are also typically ordinal. For example, since the beginning of the survey in 1972, the GSS has asked respondents "How often do you attend religious services?" The response categories are: "never," "less than once a year," "about once or twice a year," "several times a year," "about once a month," "2-3 times a month," "nearly every week," "every week," or "several times a week." These categories have an inherent rank of low to high (or less to more) religious service attendance.

Perhaps the simplest survey question is the yes/no one. This dichotomy is often used to measure past engagement with some activity or behavior. Yes/no questions are ordinal since they can be ranked. "Yes" indicates presence of some characteristic while "no" indicates its absence. For example, since 1977, the GSS has asked respondents "Were you born in this country?" As we shall see, rank-ordered response categories allow us to perform additional statistical analyses on ordinal variables than we can with nominal ones. Dichotomous variables with only two response categories that are intentionally coded as "0" versus "1" are commonly known as **dummy variables** to distinguish presence of a characteristic (1) versus its absence (0).

While the writing and preparation of the survey questions is outside of the sociologist's control when conducting secondary data analyses, both nominal and ordinal variables are typically written to be mutually exclusive and exhaustive. **Mutual exclusiveness** concerns the idea that response categories should not overlap. Each respondent should find only one of the response categories suitable. **Exhaustiveness** means that there are enough response categories to classify every respondent. All possible answers to the survey question should be available.

Scale variables are the highest level of measurement and permit the most powerful statistical analyses (many other texts refer to these as "interval-ratio," but scale is the term employed in SPSS). These are numerical

variables expressed in the original units of the variable. When survey researchers ask questions at the scale level of measurement, no response categories are necessary. Respondents provide their answers as numbers with direct meaning. The survey question, “How many years old are you?” will lead to respondents stating their age in years. Since 1975, the GSS has asked “On the average day, about how many hours do you personally watch television?” In a face-to-face or telephone survey, there would be no need for the interviewer to read out all of the possible numerical choices.

It is important to note that scale variables are single numbers and not ranges. For example, the question “What was your total family income in dollars last year?” would be a scale variable that results in a number. However, if respondents are provided with sets of income ranges and are asked to choose the response category containing their total family income, the variable must be considered ordinal. The original units and numerical metric of the variable are what give scale variables mathematical properties that are necessary to produce certain statistics.

Conclusion

This introductory chapter has attempted to provide a concise overview of the fundamentals of sociological survey research. The steps of the research process are followed by quantitative sociologists to ensure that their research methods are rigorous, objective, and replicable. Survey research is all about analyzing variables. We tend to focus upon individuals as the unit of analysis and study attributes, attitudes, and behaviors as represented in variables derived from responses to survey questions. As we learned in the levels of measurement section, there are three different types of variables: nominal, ordinal, and scale. The level of measurement of a variable determines what is mathematically possible in the analysis of it.

The next chapter introduces IBM® SPSS®, the data analysis software that is the basis of this text. The 2021 General Social Survey, a widely-used nationally representative survey, will be employed. After providing an overview of the basic functions of SPSS, we will begin our journey into descriptive statistics by covering frequency distributions and graphs.

Key Terms

Statistics, research question, unit of analysis, variable, independent variable, dependent variable, ascribed characteristics, achieved characteristics, research hypothesis, primary data, secondary data, General Social Survey, descriptive statistics, inferential statistics, response categories, values, levels of measurement, nominal, ordinal, dummy variables, mutual exclusiveness, exhaustiveness, and scale

Endnotes

1. Best, Joel. 2008. *Stat-Spotting: A Field Guide to Identifying Dubious Data*. Berkeley, CA: University of California Press.

2. Lin, Nan. 2001. *Social Capital: A Theory of Social Structure and Action*. New York, NY: Cambridge University Press.

3. McPherson, Miller, Lynn Smith-Lovin, and Matthew E. Brashers. 2006. "Social Isolation in America: Changes in Core Discussion Networks over Two Decades." *American Sociological Review* 71 (3): 353-375.

4. Linton, Ralph. 1936. *The Study of Man: An Introduction*. New York, NY: Appleton-Century-Crofts, Inc.

5. Cho, Ryan W. and Jennie E. Brand. 2019. "Life Chances and Resources" in *The Blackwell Encyclopedia of Sociology*, Ritzer, George and Chris Rojek (eds.). Hoboken, NJ: John Wiley & Sons, Ltd.

This 2025 work is licensed under CC BY-NC-SA 4.0

Quantitative Sociology Essentials: An Introduction to Survey Data Analysis using IBM® SPSS®

Ed Collom, Ph.D.
Professor of Sociology
California State University, Fullerton

Chapter 2. Inspecting Data with SPSS

With our basic understanding of the steps of the sociological research process and the levels of measurement of variables, we can now proceed. This text is based upon the widely used and user-friendly data analysis software, IBM® SPSS®. We shall primarily employ the 2021 General Social Survey, a recent version of sociology's longest-running national survey. After the overview of the major functions of SPSS, frequency distributions and graphs will be covered to provide an introduction to descriptive statistics. The chapter concludes with an overview of two frequently used SPSS commands: split file and recoding.

Chapter Objectives

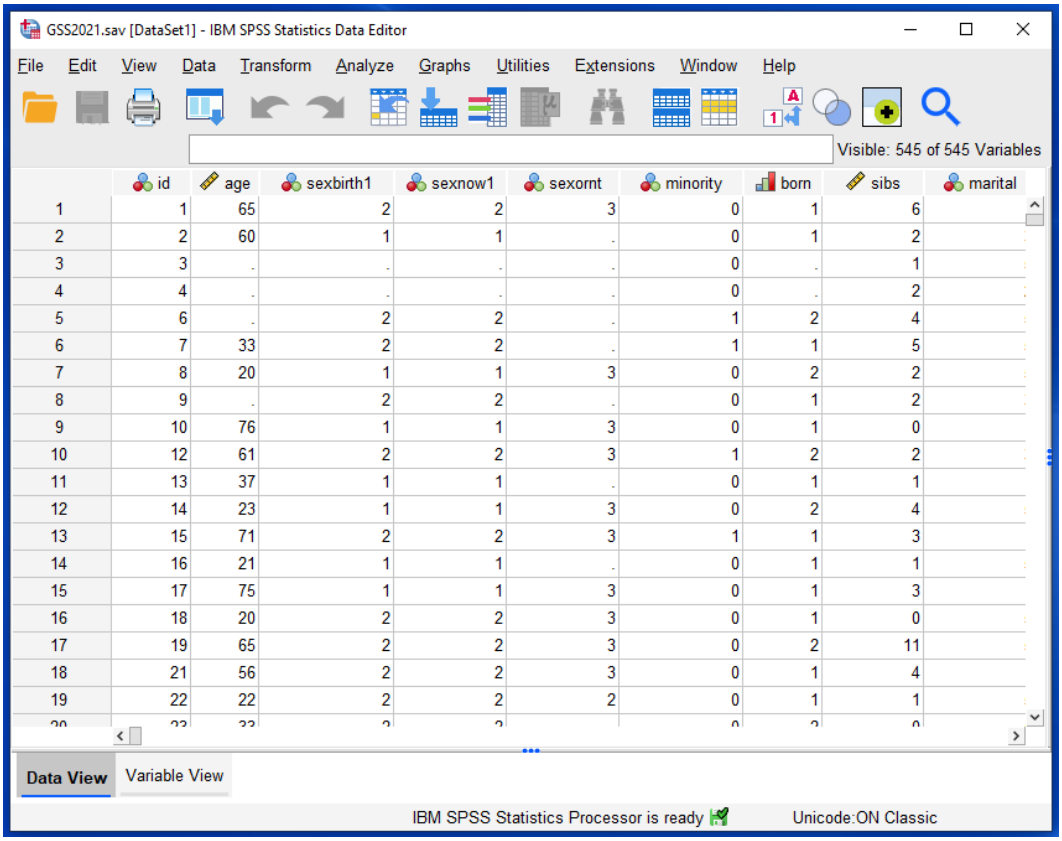
After reading this chapter, students should be able to:
- Employ SPSS to produce frequency distributions and graphs
- Report the resulting descriptive statistics from frequency distributions
- Execute the split file and recoding commands in SPSS

SPSS Basics

All Cal State Fullerton students are entitled to a free license to download and install SPSS on a personal computer (campus computers also have the software installed). SPSS uses a spreadsheet format which will be familiar to most students. The file extension for data files is .sav. The data files that will be used in this course will be located in the Canvas course website under Files > Data. The GSS2021.sav file will be employed throughout the text (open it now and follow along as you read this chapter).

Exhibit 2.1 displays the **Data View** screen upon opening the GSS 2021 datafile. This is the spreadsheet view in which the columns contain the variables and the rows contain the cases. The scrollbar at the bottom of the screen allows users to see each of the hundreds of variables. The scrollbar on the right side can be used to go to the last row of the datafile where we can see that there are a total of 4,032 cases (respondents who completed the survey).

Exhibit 2.1. SPSS Data View Screen



	id	age	sexbirth1	sexnow1	sexornt	minority	born	sibs	marital
1	1	65	2	2	3	0	1	6	
2	2	60	1	1	.	0	1	2	
3	3	.	.	.	.	0	.	1	
4	4	.	.	.	.	0	.	2	
5	6	.	2	2	.	1	2	4	
6	7	33	2	2	.	1	1	5	
7	8	20	1	1	3	0	2	2	
8	9	.	2	2	.	0	1	2	
9	10	76	1	1	3	0	1	0	
10	12	61	2	2	3	1	2	2	
11	13	37	1	1	.	0	1	1	
12	14	23	1	1	3	0	2	4	
13	15	71	2	2	3	1	1	3	
14	16	21	1	1	.	0	1	1	
15	17	75	1	1	3	0	1	3	
16	18	20	2	2	3	0	1	0	
17	19	65	2	2	3	0	2	11	
18	21	56	2	2	3	0	1	4	
19	22	22	2	2	2	0	1	1	

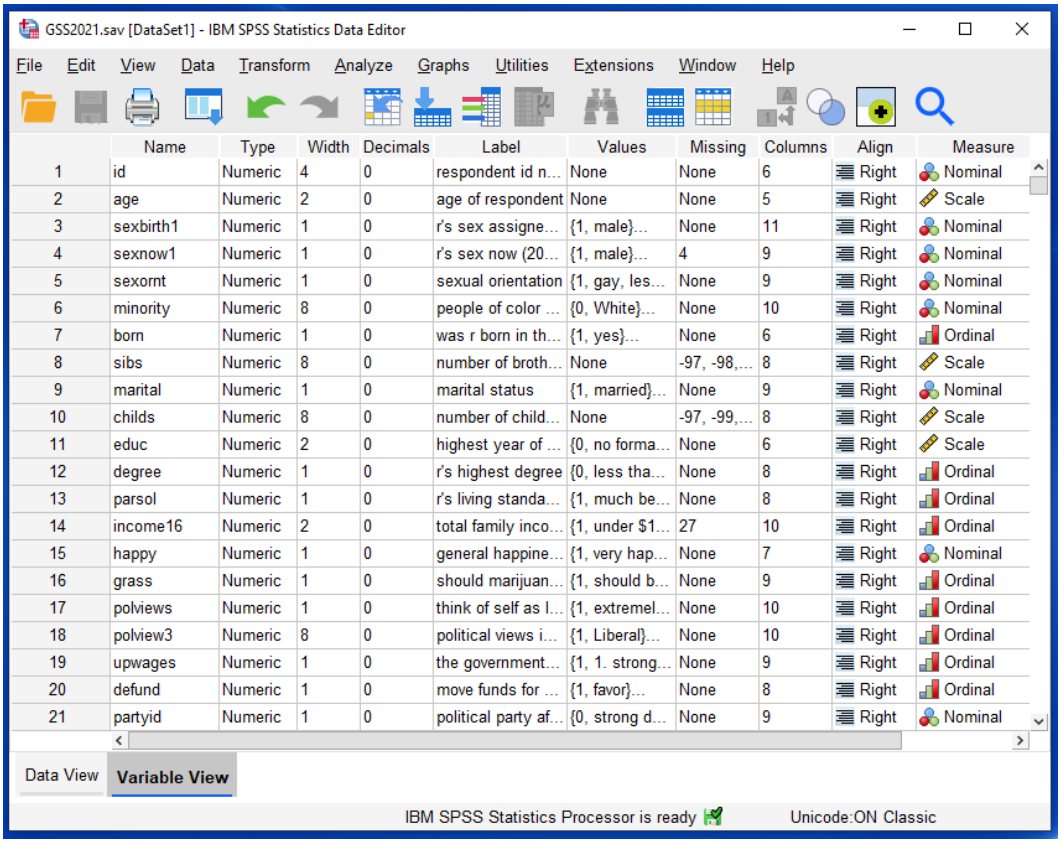
There are a number of drop-down menus along the top. In this course, we will primarily use the Analyze menu. This view is referred to as the Data View since users can see and peruse the raw data that can be analyzed in the software. The first variable is the ID variable and this is simply a unique numeric identification for each case in the dataset. The next variable is AGE, a variable that has a level of measurement of scale. The first respondent is 65 years old and the second one is 60 years old. This

view is a good reminder that as a statistical software package, SPSS only analyzes numeric information. All of the variables that we will investigate have been coded as numbers, regardless of their level of measurement.

When there is a period in the cell, it indicates that data is missing for that particular respondent on that variable. Cases 3 through 5 most likely refused to provide their age in the survey. **Missing data** is problematic since there is nothing for the sociologist to analyze. In some cases, this would reflect that the particular question was interpreted as a sensitive or intrusive one and the respondent did not feel comfortable answering it. However, there are a number of other possible explanations for missing data. A respondent may be confused by the question and not provide an answer. Alternatively, some questions are not applicable to all respondents. The GSS uses interactive filter questions in an attempt to be efficient and not bother respondents with questions that do not apply to them. For example, respondents that are not currently employed would not be asked the question about their job satisfaction. Depending upon the response to the employment status question, the job satisfaction one may automatically be skipped by the survey administration software. The GSS also uses a split-ballot design in which not all respondents are asked every question. In an effort to maximize the number of variables available to analyze, some questions are only posed to half of the respondents.

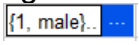
Exhibit 2.2 displays the **Variable View** screen which is found by clicking on the Variable View tab in the bottom lefthand corner. In this view, all of the variables are listed in the rows. The columns contain various information about each variable. Each variable has a Name that conforms with SPSS requirements. Some of the information in this view will not be referenced or used in this course. That is, we can ignore the Type, Width, Columns, Align, and Role columns. Most of our data will not have any Decimals, but it is possible to add them if needed. Label is an important feature as **Variable Labels** allow longer text descriptions of every variable. The survey question that is the basis of each variable is often not clear when looking at the name alone.

Exhibit 2.2. SPSS Variable View Screen



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	4	0	respondent id n...	None	None	6	Right	Nominal
2	age	Numeric	2	0	age of respondent	None	None	5	Right	Scale
3	sexbirth1	Numeric	1	0	r's sex assigne...	{1, male}...	None	11	Right	Nominal
4	sexnow1	Numeric	1	0	r's sex now (20...	{1, male}...	4	9	Right	Nominal
5	sexornt	Numeric	1	0	sexual orientation	{1, gay, les...	None	9	Right	Nominal
6	minority	Numeric	8	0	people of color ...	{0, White}...	None	10	Right	Nominal
7	born	Numeric	1	0	was r born in th...	{1, yes}...	None	6	Right	Ordinal
8	sibs	Numeric	8	0	number of broth...	None	-97, -98,...	8	Right	Scale
9	marital	Numeric	1	0	marital status	{1, married}...	None	9	Right	Nominal
10	childs	Numeric	8	0	number of child...	None	-97, -99,...	8	Right	Scale
11	educ	Numeric	2	0	highest year of ...	{0, no forma...	None	6	Right	Scale
12	degree	Numeric	1	0	r's highest degree	{0, less tha...	None	8	Right	Ordinal
13	parsol	Numeric	1	0	r's living standa...	{1, much be...	None	8	Right	Ordinal
14	income16	Numeric	2	0	total family inco...	{1, under $1...	27	10	Right	Ordinal
15	happy	Numeric	1	0	general happine...	{1, very hap...	None	7	Right	Nominal
16	grass	Numeric	1	0	should marijuan...	{1, should b...	None	9	Right	Ordinal
17	polviews	Numeric	1	0	think of self as l...	{1, extremel...	None	10	Right	Ordinal
18	polview3	Numeric	8	0	political views i...	{1, Liberal}...	None	10	Right	Ordinal
19	upwages	Numeric	1	0	the government...	{1, 1. strong...	None	9	Right	Ordinal
20	defund	Numeric	1	0	move funds for ...	{1, favor}...	None	8	Right	Ordinal
21	partyid	Numeric	1	0	political party af...	{0, strong d...	None	9	Right	Nominal

The Values field is a crucial one. As we learned in Chapter 1, nominal variables are categorical ones with response categories comprised of words that cannot be ranked. The **values** are numbers representing each of the response categories so that SPSS can provide statistical analyses of these variables. The first two variables (ID and AGE) have "None" listed in that field since these variables are already numbers with meaning. The SEXBIRTH1 variable (based on the survey question, "Was your sex recorded as male or female at birth?") does have preassigned values. After clicking on the values cell, it is highlighted and the square on the right side with the three dots turns blue when the cursor is hovered over it:



After clicking that square, the **Value Labels** dialog opens. As you will learn, SPSS uses a series of such dialogs. Here, we see that male respondents are coded "1" and female respondents are coded "2." Value labels are critical since they indicate how categorical variables have been

coded and what those values actually represent. The numeric values for nominal and ordinal variables have no inherent, direct meaning. They can be thought of as placeholders for the response category text. By opening the value labels dialog for the SEXNOW1 variable ("Do you describe yourself as male, female, or transgender?"), we see male respondents are coded "1," female respondents "2," transgender respondents "3," and those who chose "none of these" are coded as "4."

The next column under the variable view is Missing. This is where sociologists assign missing values to variables. After assigning missing values, SPSS will not analyze those particular cases. Notice that "None" appears under the missing column for the first three variables. This does not necessarily mean that these variables do not have any missing data. For example, earlier in the data view, we saw several periods appearing in the AGE variable column. Those periods are **System-missing** values that were blank cells when this dataset was imported into SPSS.

For the SEXNOW1 variable we see the value "4" in the missing column. After clicking on that cell and then the square with the three dots, the **Missing Values** dialog is opened. This dialog is for the researcher to define any missing values and is referred to as **User-missing**. In this dialog we see the value "4.000" in the first of the three boxes under Discrete missing values. This indicates that those respondents who answered "none of these" on the SEXNOW1 question will not be analyzed. It is up to the sociologist to determine which cases should be defined as user-missing. We have to think ahead about our future analyses and whether certain responses would yield meaningful information.

Sociologists and SPSS users must confirm that the variables they plan on analyzing have been defined completely and correctly prior to analyses. Most datasets need to be "cleaned" by the researcher. The GSS is no exception. The GSS2021.sav datafile does contain variable labels and value labels, but the sociologist should always check these for their accuracy. The codebook contains the exacting wording of all of the questions and can be easily referenced as we need clarification. Likewise, the GSS does have some pre-defined system and user-missing values, but there is no guarantee that all do or that all are accurate.

The last column of useful information for us in the variable view is the Measure one. This is where we define each variable's level of

measurement. Most of the variables in the GSS2021.sav datafile are listed as Nominal, the SPSS default upon importing raw data. Therefore, it is also our job to confirm that the level of measurement is defined correctly. By clicking on any of those cells, a dropdown appears for you to select Nominal, Ordinal, or Scale. Notice that each level of measurement has a unique symbol assigned to it. That symbol also appears in the data view tab before each variable name for our convenience.

SPSS users also need to learn file management skills. Throughout this term, students will be working with a variety of datasets, making changes to those datasets, and producing analyses ("output") that can also be saved. As changes are made to datasets, it is recommended to save those revisions under a new filename. The GSS2021.sav datafile should be kept in its original form so that you can go back to it if needed. Think about whether you prefer to save your files on a hard drive, flash drive, or cloud-based service such as Dropbox. If you are working on a public campus machine, you can download files to it to use, but they will need to be backed up and stored elsewhere.

Frequency Distributions

The data view in SPSS provides the raw data as lists of scores on variables. While we are able to sort that raw data by the values, it is not usually practical to use the data view in SPSS for any type of analyses. **Frequency distributions** are simple, yet powerful tools, to analyze a variable. They provide a summary of the responses and show how many cases (respondents) fall in each response category of the variable. Since they focus on only one variable at a time, frequency distributions are considered univariate analyses or descriptive statistics.

Frequency distributions are the starting point for all statistical analyses. Sociologists need to be familiar with how each variable that they are interested in is coded and how respondents are dispersed across the various categories. In addition to the number of cases in each response category, frequency distributions also provide useful percentages that standardize the data. As we shall learn, it is much easier to work with percentages than the raw numbers.

Before we run our first frequency distribution, we need to set some preferences in SPSS. The default options when the software is opened are not always the best ones for sociological survey data analysis. The

second menu item in SPSS is the Edit dropdown. After clicking Edit, the last item in the list is Options. When you activate Edit > Options there are 13 different tabs that provide a variety of different options within the software. We will only use two of these tabs to make SPSS function much more efficiently for us.

Exhibit 2.3 provides two screenshots of the General and the Output tabs. Under the General tab, you will need to click two radio buttons under the Variable Lists: “Display names” and “Alphabetical.” Variable names are much easier to work with than the labels in SPSS dialogs. Then, click the Output tab and select “Names and Labels” and “Values and Labels” under both the Outline Labeling and the Pivot Table Labeling options. Finally, click the OK button at the bottom. This directs SPSS to provide both variable names and variable labels as well as values and value labels in all of the output.

Exhibit 2.3. SPSS Edit > Options Preferences

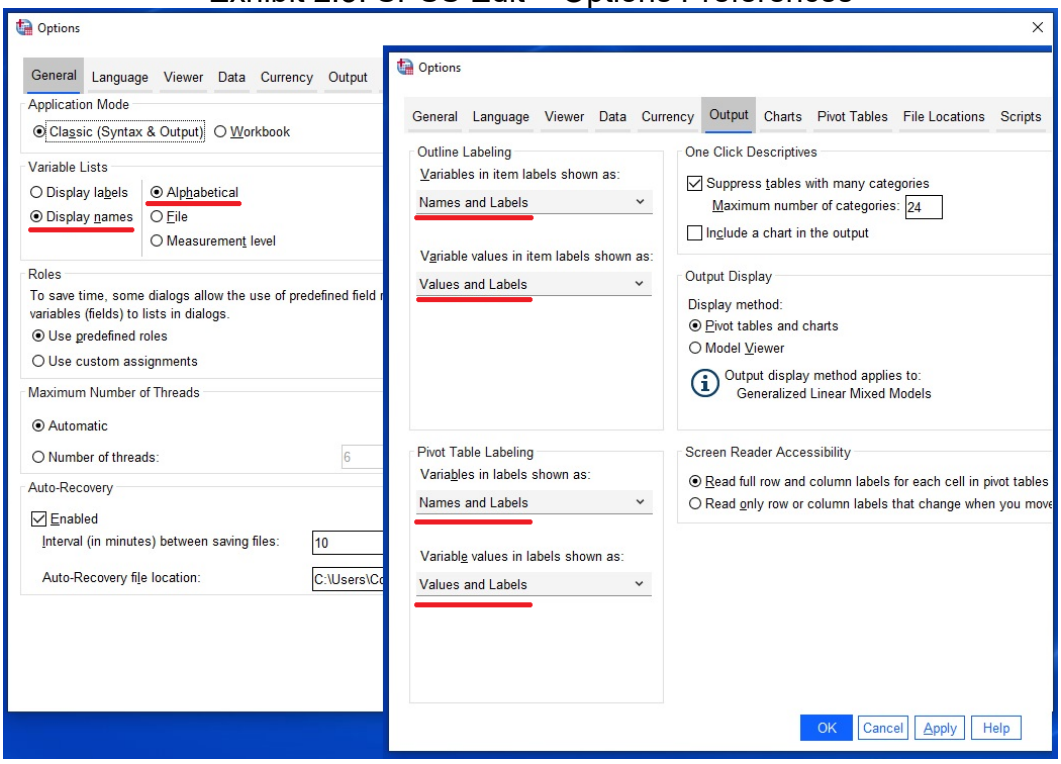
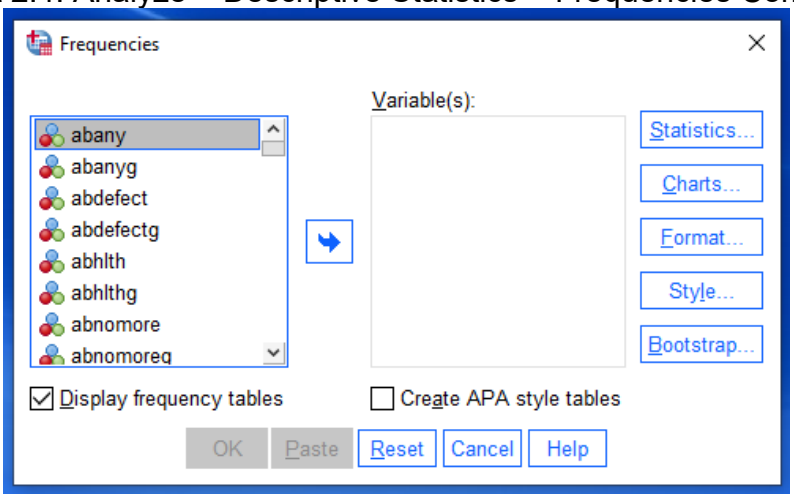


Exhibit 2.4 illustrates the dialog that is produced from the Analyze > Descriptive Statistics > Frequencies command. Notice that the list on the left shows the variable names listed in alphabetical order. Users can click on any variable in that list and then tap any letter on their keyboard and SPSS will move down to the first variable beginning with that letter.

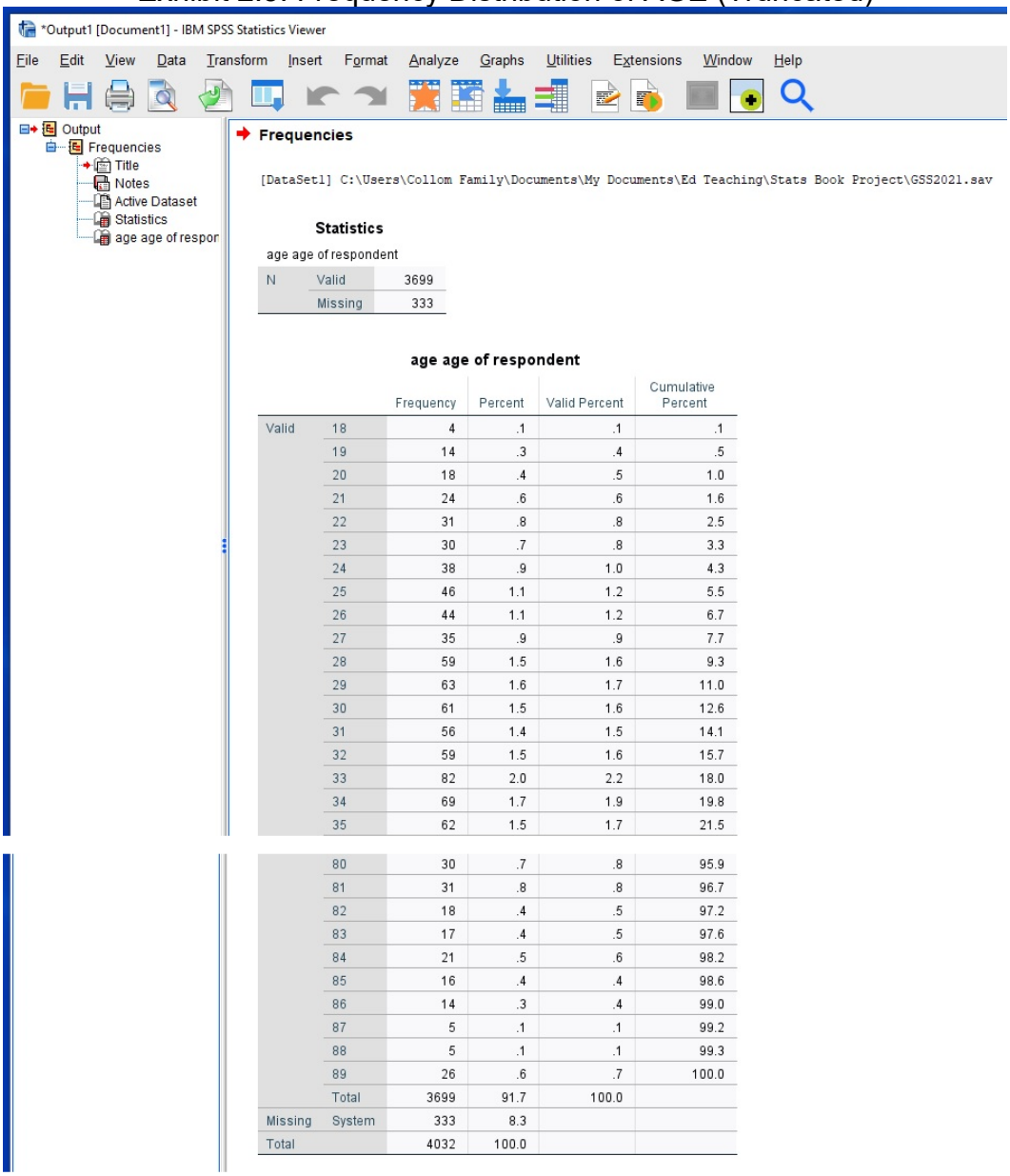
Exhibit 2.4. Analyze > Descriptive Statistics > Frequencies Command



Let's produce a frequency distribution for AGE. Scroll down to the age variable (notice that it is the first one in this list with the ruler icon symbolizing scale variables). You can double-click it to move it over to the Variable(s) box or click the arrow in the middle once it is selected. Now that there is a variable in that box, the OK button is activated. Click OK and you will have just produced your first SPSS output.

Exhibit 2.5 provides a truncated, partial version of the frequency distribution of the AGE variable from the GSS2021.sav dataset. First, you will have noticed that SPSS launched an output window. The output viewer will always be separate from the data editor window. Users can save their output as a separate file and export it to a variety of different formats if desired. The navigation pane on the left allows you to easily skip around to different parts of the output. At the top of the window, the file location of the dataset that is being used for the analyses is displayed.

Exhibit 2.5. Frequency Distribution of AGE (Truncated)



The Statistics box at the top displays valid and missing cases for the AGE variable (notice that it also shows the variable name and the variable label). This variable has 3,699 **valid responses** that we can analyze. This variable has 333 system-missing responses (as mentioned earlier, these are respondents who likely skipped this question). Notice that the

Valid and Missing cases are clearly labeled at the top and bottom of the table. The numeric values 18, 19, 20, etc. are the respondent ages in years. The column labeled Frequency shows that 4 respondents are 18 years of age, 14 are 19 years old, 18 are 20 years old, and so forth.

The next column is Percent. We rarely use this information since the percent calculation includes both valid and missing data. At the bottom of the table you can see that the 3,699 valid responses comprise 91.7% of the total cases. 8.3% of the cases (333) are missing data on this variable.

The **Valid Percent** column is the most important one in a frequency distribution. This calculation excludes the missing data and focuses upon only the valid cases. Here, we see that 0.1% of respondents are aged 18, 1.0% are 24, 2.2% are 33, and so forth.

The last column of a frequency distribution is **Cumulative Percent**. This calculation is very convenient, particularly for a variable such as AGE with so many response categories. The cumulative percent is a running total of the valid percentages. The first value will always be identical to the first valid percent (0.1% here). Then, the next valid percent (0.4% for those aged 19) is added to the previous one to get 0.5%. This indicates that 0.5% of the respondents are aged 18 or 19. Working down the table, we see that 11.0% of the respondents are 29 or younger and 21.5% are 35 or younger. The cumulative percent will always add up to be 100.0% in the final response category (89 here).

Now, let's run a frequency distribution for the SEXBIRTH1 variable (a nominal one). Go back to the Analyze > Descriptive Statistics > Frequencies command. You will notice that AGE remains in the Variable(s) list. SPSS automatically saves the previous command in each dialog. You will need to get in the habit of clicking the Reset button in the middle at the bottom of the dialog to clear out the previous variable(s) once you are ready to move on. Since the variables are listed by name alphabetically, hit "S" on your keyboard, scroll down to the SEXBIRTH1 variable, double-click it, and hit OK.

Exhibit 2.6. Frequency Distribution of SEXBIRTH1

Frequencies

Statistics

sexbirth1 r's sex assigned at birth (2021)

N	Valid	3928
	Missing	104

sexbirth1 r's sex assigned at birth (2021)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 male	1730	42.9	44.0	44.0
	2 female	2198	54.5	56.0	100.0
	Total	3928	97.4	100.0	
Missing	System	104	2.6		
Total		4032	100.0		

As seen in Exhibit 2.6, the SEXBIRTH1 variable has 3,928 valid cases and 104 missing cases. 44.0% of the respondents were assigned the male sex at birth and 56.0% female. You may have expected these percentages to be closer to one another since they actually are in the U.S. population. These results show that more females completed the 2021 General Social Survey than males (Census data from 2021 indicates that around 49.0% of the U.S. population was male and 51.0% female). The differences between samples and populations (and potential discrepancies) will be covered in Chapter 5.

Finally, let's run a frequency distribution for ATTEND, an ordinal variable. As we learned in Chapter 1, the ATTEND variable (Exhibit 2.7) is the frequency of the respondent's religious service attendance. Here, we see that 29.7% of respondents report "never" attending religious services and 14.3% attend "less than once a year." As we saw with AGE, the cumulative percent is useful in navigating frequency distributions with many response categories. It is also helpful in reporting results for ordinal ones. According to the cumulative percent, 55.4% of respondents report attending "about once or twice a year" or "less than once a year" or "never." In other words, the majority of respondents (more than half)

seldomly attend religious services. These types of statements are possible since this is an ordinal variable that is ranked from low to high attendance. Also notice that the values of the response categories are coded from "0" to "8." Zero implies the absence of some characteristic. So, it makes since that the GSS sociologists decided to code "never" as zero.

Exhibit 2.7. Frequency Distribution of ATTEND

Frequencies

Statistics

attend how often r attends religious services

N	Valid	3962
	Missing	70

attend how often r attends religious services

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 never	1178	29.2	29.7	29.7
	1 less than once a year	565	14.0	14.3	44.0
	2 about once or twice a year	453	11.2	11.4	55.4
	3 several times a year	403	10.0	10.2	65.6
	4 about once a month	122	3.0	3.1	68.7
	5 2-3 times a month	200	5.0	5.0	73.7
	6 nearly every week	331	8.2	8.4	82.1
	7 every week	532	13.2	13.4	95.5
	8 several times a week	178	4.4	4.5	100.0
	Total	3962	98.3	100.0	
Missing	System	70	1.7		
Total		4032	100.0		

As evident, frequency distributions are very powerful and useful. They are always the starting point of our analyses since they show the response categories, how they are coded (their values), the number of valid and missing cases, and the valid percent of respondents within each category

of the variable. Frequency distributions are often complemented with graphs to visualize the survey responses.

Graphs

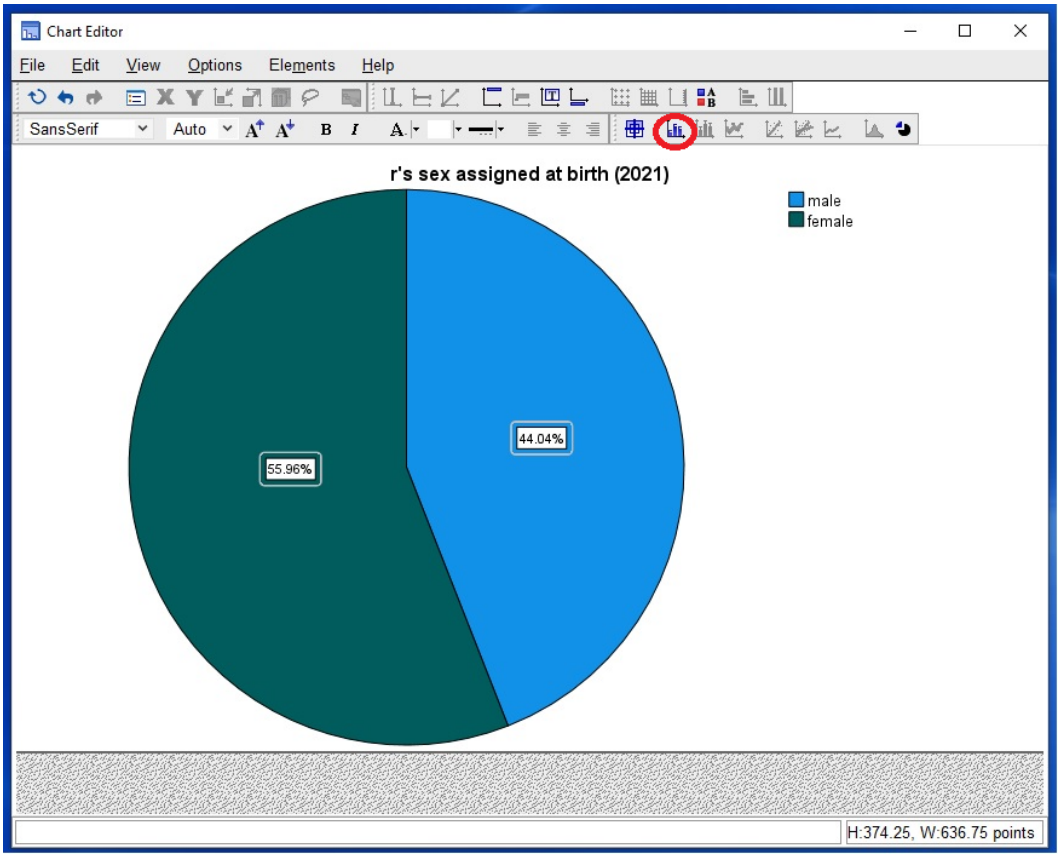
As we shall see throughout this text, a key job of the quantitative sociologist is to choose the appropriate statistical technique for the analysis. When it comes to graphs, the level of measurement of the variable is key. **Pie charts** should be used for nominal variables. **Bar charts** should be used for ordinal variables. And, **histograms** should be used for scale variables.

The easiest way to produce graphs is through the Analyze > Descriptive Statistics > Frequencies command that we just learned. In that dialog, the second button down on the right side is labeled Charts. Let's create charts for the frequency distributions we investigated earlier. We'll begin with the nominal variable, SEXBIRTH1, and create our first pie chart. Move SEXBIRTH1 into the Variable(s) box, click Charts, select the Pie chart radio button, select the Percentages radio button under Chart Values, click Continue, and then click OK. You may notice that SPSS takes longer to produce graphs than tables.

Once the pie chart appears in the output window, double-click it to launch the Chart Editor window. Now, click on the Show Data Labels icon circled in Exhibit 2.8 (the symbol looks like a bar chart) and the valid percentages appear within the two pieces of the pie (you can also use the Elements > Show Data Labels command if desired). Click Close on the Properties window that appears and then close the Chart Editor by clicking on the "X" in the upper, right-hand corner of the window. Now, your completed pie chart appears in the output window. Notice that the title is listed as the variable label and the legend containing the value labels and corresponding color is automatically included.

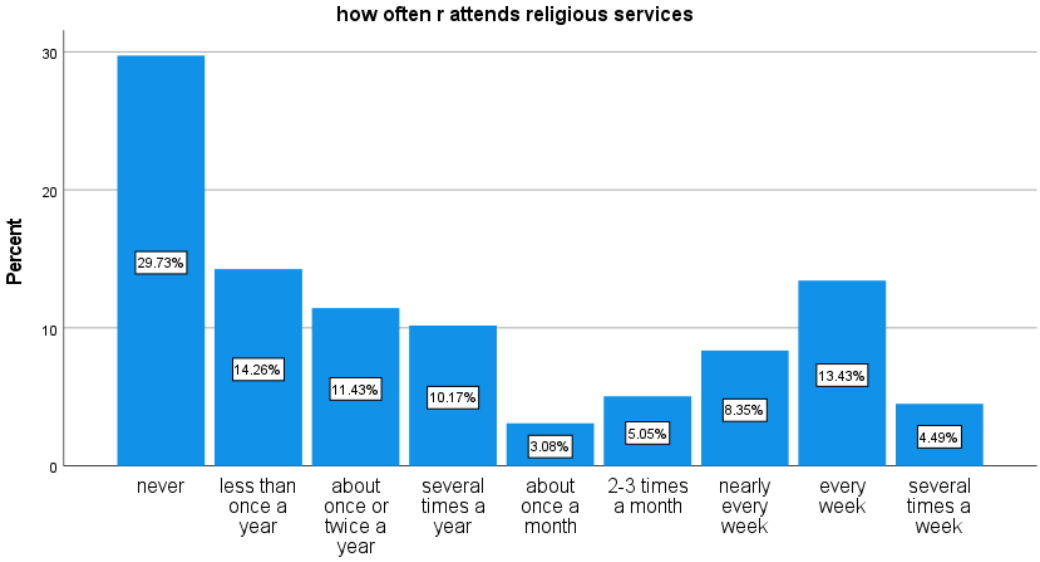
This pie chart contains the most important portion of the frequency distribution, the valid percent. Graphs complement frequency distributions by providing a visual graphic that is easy to understand.

Exhibit 2.8. Chart Editor Feature for Pie Chart of SEXBIRTH1



Now, let's create a bar chart for the ordinal variable ATTEND. Go to Analyze > Descriptive Statistics > Frequencies and hit Reset. Move ATTEND into the box, select Bar charts and Percentages from the Charts dialog, hit Continue, and then OK. Double-click the bar chart to activate the Chart Editor (or, on a PC you can use the right-click button and select Edit), click the Show Data Labels icon, close the Properties window, and then close the Chart Editor. You should see something very similar to Exhibit 2.9.

Exhibit 2.9. Bar Chart of ATTEND



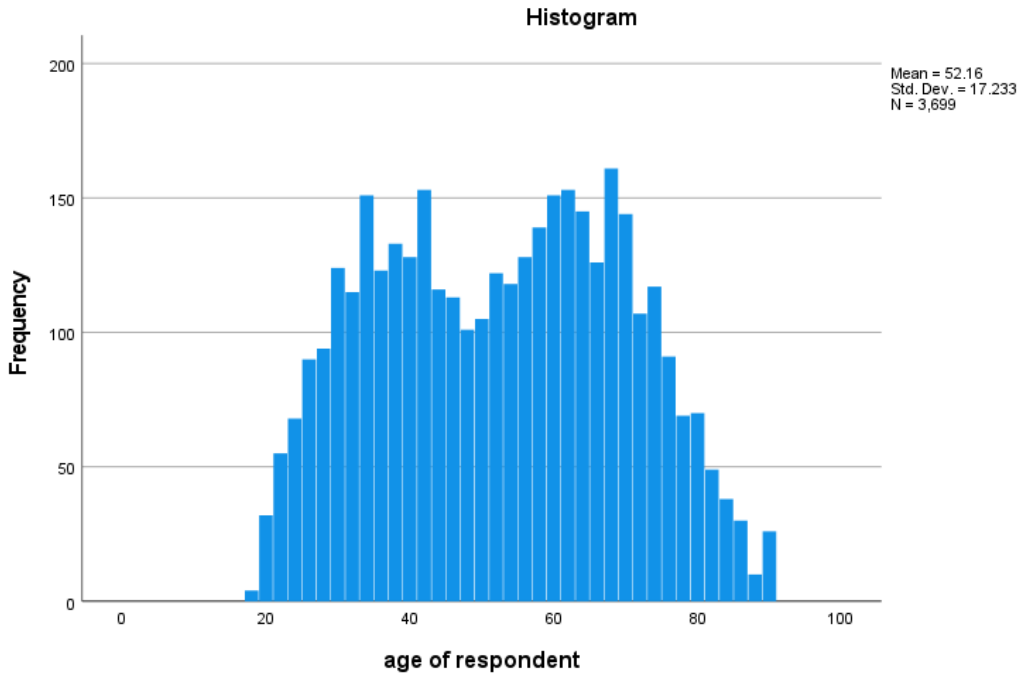
Notice that the vertical, Y-axis of the chart is labeled Percent since we requested percentages instead of frequencies. The valid percentages within each category appear in their respective bars per our Show Data Labels command.

Bar charts are particularly good at helping us determine how dispersed the cases are across the various response categories. We'll spend the entire Chapter 4 on measures of dispersion, but for now we can think about **heterogeneity** and **homogeneity**. Heterogeneity refers to differences or variation. Heterogeneous distributions are ones in which the respondents are fairly evenly dispersed or distributed across all of the response categories. In a perfectly heterogeneous distribution, the height of each of the bars would be the same (and the valid percentages would all be equal). This would indicate that there is maximum variation or differences on this variable.

Homogeneity refers to sameness or similarities. Homogeneous distributions are those in which respondents are concentrated or clustered into just a few categories of the variable. A perfectly homogeneous distribution would have 100% of the respondents in one category and 0% in all of the others. In other words, the respondents would be the same on this variable.

Histograms are the appropriate graph for variables measured at the scale level. So, let's produce a histogram for AGE. Histograms only use frequencies on the Y-axis and we will not employ the Show Data Labels command since it makes the chart too busy. As seen in Exhibit 2.10, the bars of a histogram are contiguous (they touch each other) to represent the numeric nature of scale variables.

Exhibit 2.10. Histogram of AGE



As you may notice, the graph contains only about half of the number of bars (37) as there are categories in the AGE variable (72). SPSS automatically combines the categories of a variable when producing a histogram to avoid making the graph too busy and to demonstrate the overall shape of the distribution. While we will not explore these features here, users can change the options to specify the number of bars (or "bins") and/or the width of each.

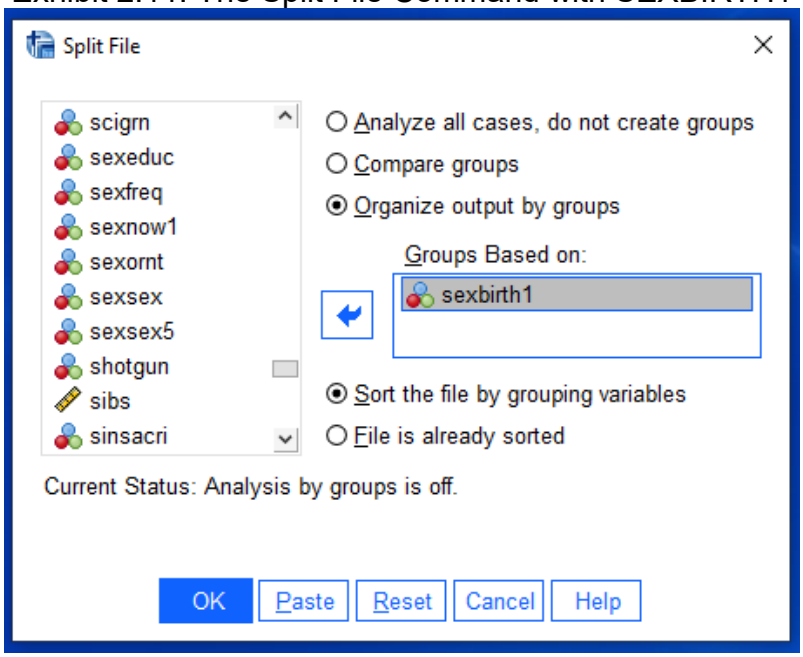
Histograms are often used in conjunction with the normal curve, a concept we will learn about in Chapter 5. We can also see that there are some statistics included with the graph. The mean will be covered in the next

chapter and the standard deviation in Chapter 4. For now, the graph shows us that there is some variety or heterogeneity in the responses. We can also see a “valley” in the middle dividing the peaks on either side.

Split File Command

The SPSS **Split File** command is frequently used by sociologists and survey researchers in general. The command simply splits any output that we produce into groups. Click on the Data menu and then Split File (third from the bottom). We are going to split our file by SEXBIRTH1.

Exhibit 2.11. The Split File Command with SEXBIRTH1



The Split File dialog should be completed as seen in Exhibit 2.11. Click the Organize output by groups radio button, move SEXBIRTH1 into the box, click Sort the file by grouping variables, and then click OK. When you navigate back to the data editor window, you will notice that the bottom, right-hand corner of the window states: Split by sexbirth1. This means that all of the output you now request will be separated by males and females. This allows us to see if these groups differ from one another on variables of interest. We will continue to investigate frequency

distributions for now. Use the Analyze > Descriptive Statistics > Frequencies command and select POLVIEWS as the variable.

The POLVIEWS variable is an ordinal one based on the following GSS survey question: “We hear a lot of talk these days about liberals and conservatives. I’m going to show you a seven-point scale on which the political views that people might hold are arranged from extremely liberal, point 1, to extremely conservative, point 7. Where would you place yourself on this scale?”

Exhibit 2.12 shows most of the output that is produced (note that SPSS does provide output for those respondents with missing data on the SEXBIRTH1 variable, but we can skip that since it does not have any relevance). The frequency distribution for the male respondents is at the top and the females are at the bottom. We can see in the Cumulative Percent that 31.4% of males identify as either (“extremely liberal” or “liberal” or “slightly liberal”). For females, 35.3% identify as some kind of liberal. Thus, females are slightly more likely to identify as liberal. 32.2% of males and 36.3% of females identify as moderate.

We can also sum up the valid percentages of males and females who identify as some kind of conservative. For males, it is 36.5% (14.4 + 17.7 + 4.4) and for females it is 28.5% (10.3 + 13.9 + 4.3). Thus, males are slightly more likely to identify as conservative.

The Split File command is a useful one to start thinking about how two variables may be associated with one another. The key thing about using this command is to remember to turn it off! Go back to Data > Split File, select the first radio button “Analyze all cases, do not create groups,” and click OK. You will notice that the warning in the bottom, right-hand corner of the data editor window has now disappeared.

Exhibit 2.12. POLVIEWS Frequency Distributions Split by SEXBIRTH1

polviews think of self as liberal or conservative[a]

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 extremely liberal	87	5.0	5.1	5.1
	2 liberal	253	14.6	14.7	19.8
	3 slightly liberal	198	11.4	11.5	31.4
	4 moderate, middle of the road	552	31.9	32.2	63.5
	5 slightly conservative	247	14.3	14.4	77.9
	6 conservative	303	17.5	17.7	95.6
	7 extremely conservative	76	4.4	4.4	100.0
	Total	1716	99.2	100.0	
Missing	System	14	.8		
Total		1730	100.0		

a. sexbirth1 r's sex assigned at birth (2021) = 1 male

polviews think of self as liberal or conservative[a]

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 extremely liberal	116	5.3	5.4	5.4
	2 liberal	362	16.5	16.7	22.1
	3 slightly liberal	285	13.0	13.2	35.3
	4 moderate, middle of the road	784	35.7	36.3	71.6
	5 slightly conservative	223	10.1	10.3	81.9
	6 conservative	300	13.6	13.9	95.7
	7 extremely conservative	92	4.2	4.3	100.0
	Total	2162	98.4	100.0	
Missing	System	36	1.6		
Total		2198	100.0		

a. sexbirth1 r's sex assigned at birth (2021) = 2 female

The Recode Command

Another commonly used SPSS feature is the **recode** command. We use this command on variables with many categories to collapse them into fewer ones. Sometimes there are too few respondents in particular categories of a variable for certain statistical analyses. In these instances, when it makes sociological sense, we can combine some of the categories. Other times, we may have a scale variable with many response categories when we would prefer to simplify it into an ordinal variable with fewer categories comprised of range intervals.

Remember the ATTEND variable from Exhibits 2.7 and 2.9? It has nine response categories and may not be feasible to use in some statistical analyses such as crosstabulations (the topic of Chapter 8). So, let's recode it into just three categories representing low, moderate, and high religious service attendance.

The command that we will employ is Transform > Recode into Different Variables. This will create a new variable whereas the Recode into Same Variables is rarely used since it overwrites the original data in the existing variable. The Recode into Different Variables dialog requires the Input Variable to be identified (that is, the original source variable) and the new Output Variable to be named (and optionally provided with a Variable Label). Select the ATTEND variable and move it into the box. Type the new variable name, ATTEND3, into the Name field of the Output Variable box. You can also add a label such as "attend recoded into 3 cats." Then, click the Change button and you will see the question mark in the main box change to ATTEND3.

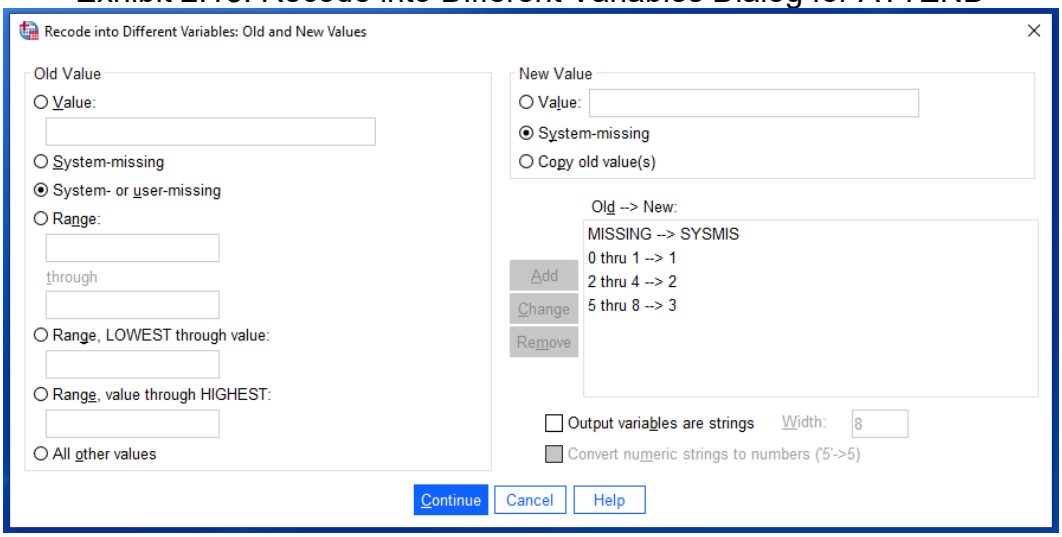
Next, click the Old and New Values button. The Recode into Different Variables: Old and New Values dialog requires the user to specify how the old values of the ATTEND variable will be recoded into the new version of the variable, ATTEND3. Let's consider the old values "never" (0) and "less than once a year" (1) to be "low" attendance. The old values "about once or twice a year" (2), "several times a year" (3), and "about once a month" (4) to be "moderate" attendance. And, the old values "2-3 times a month" (5), "nearly every week" (6), "every week" (7), and "several times a week" (8) to be "high" attendance.

We will use the Range feature on the Old Value (left side of the) dialog. Click the Range radio button and type "0" into the first box and "1" into the second box. This tells SPSS to take the old values of 0 and 1 and

combine them into a new variable category. We will code that "low" category as 1, so type "1" into the Value box on the right side under New Value. Once this is done, you will notice that the Add button becomes active. Click that now and you will see the syntax "0 thru 1 --> 1" in the Old --> New box.

Continue by recoding the old values 2 through 4 into the new value of 2 ("moderate") and the old values 5 though 8 into the new value of 3 ("high"). The last thing that we always need to do when recoding is to make sure that any missing values in the old variable remain as missing in the new variable. Thus, click the third radio button (System- or user-missing) on the left, the second radio button (System-missing) on the right, and then Add. The dialog should now appear as it is in Exhibit 2.13.

Exhibit 2.13. Recode into Different Variables Dialog for ATTEND



Click the blue Continue button and then click OK on the first dialog. You have just created your first recoded variable! Now, let's clean it up a bit. Click on the Variable View tab of the data editor screen. Scroll all the way to the end and the new ATTEND3 variable will be listed. Click on the Decimals box and then click the down arrow symbol twice to remove the decimal places in this new variable (they are not relevant). Then, click the Values box where it says None. Now, click the square with the three dots. This launches the Value Labels dialog. We need to label our values so

that we can see the new categories when analyzing this variable. Otherwise, we will only see the values “1,” “2,” and “3.”

Type “1” into the Value box, type “low” into the Label box, and click +. Type “2” into the Value box, type “moderate” into the Label box, and click +. Finally, type “3” into the Value box, type “high” into the Label box, and then click OK. Now you will see that the “None” in the Values cell has been replaced with the new value labels. The last step is to identify the level of measurement of this new variable. Scroll over to the Measure column, click the dropdown, and select Ordinal. Now that you have made edits to your datafile, it is a good time to save it. Use the File > Save As… command. Change the File name to something else such as “My GSS2021.sav,” choose where you want to save it up at the top “Look in:” dropdown, and click Save.

Exhibit 2.14. Frequency Distribution for ATTEND3

Frequencies

Statistics

attend3 attend recoded into 3 cats

N	Valid	3962
	Missing	70

attend3 attend recoded into 3 cats

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 low	1743	43.2	44.0	44.0
	2 moderate	978	24.3	24.7	68.7
	3 high	1241	30.8	31.3	100.0
	Total	3962	98.3	100.0	
Missing	System	70	1.7		
Total		4032	100.0		

Now, confirm that your recode work is accurate. Use the Analyze > Descriptive Statistics > Frequencies command and generate a frequency distribution for ATTEND3. Refer to Exhibit 2.14 to confirm that your output

matches it. Notice that there are 70 missing cases here as in the original ATTEND variable. 44.0% of the respondent are in the “low” religious service attendance category and this matches the Cumulative Percent of the “less than once a year” category in the original ATTEND variable.

SPSS user skills will only be developed by actively using the software on a regular basis. So, feel free to practice by recoding another. Remember that SPSS automatically saves the last command used within each dialog. When recoding multiple variables, it is critical to hit the Reset button to clear out the previous command. Otherwise, your next recoded variable will likely have errors.

As seen in Exhibit 2.5, AGE has 72 valid categories ranging from 18 to 89 years. Recode the AGE scale variable into a new ordinal variable entitled AGECATS. The new variable can have five age categories: 18-29, 30-44, 45-59, 60-74, and 75-89 years old. Confirm that your recode is accurate by comparing it to Exhibit 2.15.

Exhibit 2.15. Frequency Distribution for AGECATS

Frequencies

Statistics

agecats age recoded into 5 cats

N	Valid	3699
	Missing	333

agecats age recoded into 5 cats

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 18-29	406	10.1	11.0	11.0
	2 30-44	980	24.3	26.5	37.5
	3 45-59	907	22.5	24.5	62.0
	4 60-74	1023	25.4	27.7	89.6
	5 75 and older	383	9.5	10.4	100.0
	Total	3699	91.7	100.0	
Missing	System	333	8.3		
Total		4032	100.0		

This chapter has provided an introduction to SPSS basics using the 2021 General Social Survey. We then learned about frequency distributions and graphs which are descriptive statistics enabling us to describe our data. The chapter concluded with step-by-step tutorials on using the split file and recode commands in SPSS. In Chapter 3, we continue our journey into descriptive, univariate statistics with measures of central tendency.

Key Terms

SPSS data view, missing data, SPSS variable view, variable labels, values, value labels, system-missing, missing values, user-missing, frequency distributions, valid responses, valid percent, cumulative percent, pie charts, bar charts, histograms, heterogeneity, homogeneity, split file, and recode

This 2025 work is licensed under CC BY-NC-SA 4.0

Quantitative Sociology Essentials:
An Introduction to Survey Data Analysis using IBM® SPSS®

Ed Collom, Ph.D.
Professor of Sociology
California State University, Fullerton

Chapter 3. Measures of Central Tendency

Measures of central tendency assist sociologists in finding patterns in the survey data they are analyzing. These are descriptive or "univariate" statistics since they are applied to only one variable at a time. They provide a single summary statistic describing the average or typical score in a distribution. Sociologists choose the appropriate measure of central tendency according to the level of measurement of the variable. In this chapter, we will review the characteristics of the mode, median, and mean, learn when to employ which one, and use SPSS to produce these statistics on variables from the 2021 General Social Survey.

Chapter Objectives

After reading this chapter, students should be able to:
- Select and produce appropriate measures of central tendency in SPSS
- Report values of measures of central tendency and describe distributions
- Compare values of measures of central tendency across groups and draw conclusions

Mode

The **mode** is the most common score among the valid responses of a frequency distribution. It is the response category containing the highest frequency of valid cases. The mode is the only measure of central tendency that is to be used with nominal variables. Recall that nominal variables are comprised of text-based response categories that cannot be ranked.

The statistic itself is always the value of the response category (how it is coded in SPSS) with the largest number of cases, not the actual number of respondents in that particular category. Since the response category values of nominal variables are mere placeholders (they do not have any inherent meaning), researchers must also report the value label (or meaning) for the modal category.

Quantitative sociologists always have to be very careful in reporting results. Note that while the mode is the most frequently occurring response in a frequency distribution, it is not necessarily the response category in which most respondents fall. When the term **most** is used to describe respondents to a survey, it has the meaning of majority (more than half). So, be most careful when using the term most.

As discussed in the first chapter, the 2021 GSS collected sex and gender identity in two separate questions. We already investigated the frequency distribution of the former, SEXBIRTH1, in Exhibit 2.6. So, let's begin by focusing on SEXNOW1, current gender identity: "Do you describe yourself as male, female, or transgender?" This is a nominal variable since the response categories cannot be ranked. Thus, the mode is the appropriate measure of central tendency for SEXNOW1.

Measures of central tendency are found within the frequency distribution dialog in SPSS. Open GSS2021.sav, use the Analyze > Descriptive Statistics > Frequencies command, and move SEXNOW1 into the Variable(s) box. Now, click on the Statistics box in the upper-right corner. Place a checkmark next to Mode in the Central Tendency box, hit Continue, and then click OK.

Exhibit 3.1 contains the SPSS output. The statistics box reports that there are 3,916 valid cases and 116 missing cases on this variable. The mode is reported in the last row of that table as 2. The frequency distribution indicates that 44.1% of respondents identify as male, 55.6% as female, and 0.3% as transgender. The most common respondent, and the majority of cases, are female. We could report this statistic as follows:

> The SEXNOW1 variable is a nominal one, so the mode is the appropriate measure of central tendency. The most common score is 2, which represents female respondents. Females also comprise the majority (55.6%) of respondents.

Exhibit 3.1. Mode and Frequency Distribution for SEXNOW1

Frequencies

Statistics

sexnow1 r's sex now (2021)

N	Valid	3916
	Missing	116
Mode		2

sexnow1 r's sex now (2021)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 male	1727	42.8	44.1	44.1
	2 female	2179	54.0	55.6	99.7
	3 transgender	10	.2	.3	100.0
	Total	3916	97.1	100.0	
Missing	4 none of these	16	.4		
	System	100	2.5		
	Total	116	2.9		
Total		4032	100.0		

While it is a rarity with large datasets such as the GSS, the SPSS output can indicate that there are multiple modes. This would be the instance in which two valid response categories contain the exact same number of cases and that number is greater than the frequency found in any other valid response category of the distribution.

Let's consider marital status next. The MARITAL variable is based on the question: "Are you currently married, widowed, divorced, separated, or have you never been married?" This is also a nominal variable since the response categories are text-based and cannot be ranked.

As seen in Exhibit 3.2, the mode is 1 which represents married respondents. In this case, married respondents do not quite comprise a majority: 49.7% of respondents report being married.

Now, let's test our skills and use the split file command to see if marital status varies by gender identity. Use the Data > Split file command (refer

to Chapter 2 if needed) and organize the output by SEXNOW1 groups. Rerun the MARITAL frequency distribution. Does the mode vary across the three groups? Confirm for yourself that the answer is no. The mode for male, female, and transgender respondents are all the same at 1 (married). Do not forget to turn off the split file command now as we proceed!

Exhibit 3.2. Mode and Frequency Distribution for MARITAL

Frequencies

Statistics

marital marital status

N	Valid	4023
	Missing	9
Mode		1

marital marital status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 married	1999	49.6	49.7	49.7
	2 widowed	301	7.5	7.5	57.2
	3 divorced	655	16.2	16.3	73.5
	4 separated	96	2.4	2.4	75.8
	5 never married	972	24.1	24.2	100.0
	Total	4023	99.8	100.0	
Missing	System	9	.2		
Total		4032	100.0		

Median

The **median** is the middle score in a distribution that divides it into two equal parts. It is the appropriate measure of central tendency that is to be used with ordinal variables. Recall that ordinal variables are comprised of text-based response categories that are ranked. In Chapter 2, we generated a frequency distribution for the ATTEND variable (Exhibit 2.7). It ranges from “never” which is coded 0 to “several times a week” which is coded 8. If we listed out all of the 0 responses, all of the 1 responses, and

so forth, the median would be the response associated with the middle case.

Fortunately, we do not have create tedious lists since SPSS calculates the median for us. Additionally, the median is easy to spot in a frequency distribution of an ordinal variable. Remember that the cumulative percent in a frequency distribution is the running total. By definition, the median is the 50th **percentile** of a distribution. Thus, the response category that contains the cumulative percent of 50% is the median. Since the response category values of ordinal variables are just placeholders (they do not have any inherent meaning), researchers must also report the value label (or meaning) for the median category.

Exhibit 3.3. Median and Frequency Distribution for ATTEND

Frequencies

Statistics

attend how often r attends religious services

N	Valid	3962
	Missing	70
Median		2.00

attend how often r attends religious services

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 never	1178	29.2	29.7	29.7
	1 less than once a year	565	14.0	14.3	44.0
	2 about once or twice a year	453	11.2	11.4	55.4
	3 several times a year	403	10.0	10.2	65.6
	4 about once a month	122	3.0	3.1	68.7
	5 2-3 times a month	200	5.0	5.0	73.7
	6 nearly every week	331	8.2	8.4	82.1
	7 every week	532	13.2	13.4	95.5
	8 several times a week	178	4.4	4.5	100.0
	Total	3962	98.3	100.0	
Missing	System	70	1.7		
Total		4032	100.0		

Let's run a frequency distribution for ATTEND again, but this time, request the median from the statistics dialog. As seen in Exhibit 3.3, the median is reported as 2.00, the value representing "about once or twice a year." Also, notice that the cumulative percent for "less than once a year" (1) is 44.0% and for "about once or twice a year" (2) it is 55.4%. Thus, we can clearly see that the 50th percentile is a response within that "about once or twice a year" (2) category.

Let's look at another example of the median. DEGREE is a variable that identifies the highest educational degree that a respondent has earned. As seen in Exhibit 3.4, it ranges from 0 (those who have not graduated from high school) to 4 (those with graduate school degrees). The median is reported as 2.00. The middle case is a respondent with an Associate's or junior college degree. Again, note the cumulative percent and how the 50th percentile also falls in this category.

Exhibit 3.4. Median and Frequency Distribution for DEGREE

Frequencies

Statistics

degree r's highest degree

N	Valid	4009
	Missing	23
Median		2.00

degree r's highest degree

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 less than high school	246	6.1	6.1	6.1
	1 high school	1597	39.6	39.8	46.0
	2 associate/junior college	370	9.2	9.2	55.2
	3 bachelor's	1036	25.7	25.8	81.0
	4 graduate	760	18.8	19.0	100.0
	Total	4009	99.4	100.0	
Missing	System	23	.6		
Total		4032	100.0		

Since we will use the DEGREE variable regularly in this text, it is important to remember that the GSS is a sample survey of American adults aged 18

and older. Very few respondents are traditional high school or college age young people (see Exhibit 2.5.: only 2.5% of all respondents are aged 18 to 22). DEGREE is not about student status. Respondents reporting that they have not graduated from high school are mostly older people who never earned their high school diploma.

Overall, the median is a useful snapshot of the central score of a distribution. As we shall soon see, it is also useful in determining the shape of a distribution.

Now, let's split our file by MARITAL to determine if religious service attendance is different among the various marital statuses. As you will see, SPSS creates a lot of output for this one. Exhibit 3.5 only displays the statistics boxes with the median scores for each group.

Exhibit 3.5. Median for ATTEND by MARITAL Groups

Statistics[a]

attend how often r attends religious services

N	Valid	1964
	Missing	35
Median		2.00

a. marital marital status = 1 married

Statistics[a]

attend how often r attends religious services

N	Valid	296
	Missing	5
Median		4.00

a. marital marital status = 2 widowed

Statistics[a]

attend how often r attends religious services

N	Valid	644
	Missing	11
Median		2.00

a. marital marital status = 3 divorced

Statistics[a]

attend how often r attends religious services

N	Valid	92
	Missing	4
Median		3.00

a. marital marital status = 4 separated

Statistics[a]

attend how often r attends religious services

N	Valid	957
	Missing	15
Median		1.00

a. marital marital status = 5 never married

We do see different values of the median of ATTEND for the various marital statuses. Those who have never been married (group 5 on MARITAL) have a median of 1.00 ("less than once a year"). Those who are married (1) and those who are divorced (3) both have a median of 2.00 ("about once or twice a year"). Those who are separated (4) have a median of 3.00 ("several times a year"). Finally, those who are widowed (2) have a median of 4.00 ("about once a month").

Overall, these differing median values indicate that there appears to be some association between marital status and religious service attendance.

Those who have never been married attend services less frequently than those who are widowed. Why is this? Let's think about that seventh step of the research process from Chapter 1. How can we interpret this finding? The never-married group would be comprised of single people as well as cohabitating couples. Robert Wuthnow is a leading sociologist of religion and has studied this very issue.[1] Participation in congregations is aimed more at families than individuals and attendance by single people has declined over time. Also, some cohabitators are unconventional people as they choose not to support the institution of marriage. They may also not support the institution of religion and its practices. Widowed people may have attended services with their spouses in the past and continue to go for social and spiritual reasons. It is likely that age is playing a role here as well. Those who have never been married are younger people on average and those who are widowed are likely to be older on average. We can investigate the association between ATTEND and AGE in the next section!

When using the split file command, do not get confused about the level of measurement and which measure of central tendency to employ. Only the level of measurement of the main variable of interest matters. We know that ATTEND is ordinal, so we chose the median. The split file command simply groups the output by the categories of some other variable. Its level of measurement does not change the fact that we are using the median for ATTEND. For now, turn off that split file.

Mean

Most readers will be familiar with the **mean** or average. Perhaps you create a budget and have calculated your average expenses each month for groceries, utilities, and such. The mean is a simple calculation: add up all of the scores in a distribution and then divide by the number of scores. The mean is the appropriate measure of central tendency for scale variables. Recall that these are numerical variables expressed in their original units. These are the survey questions that respondents answer by replying with a number. There is only a statistic when working with the mean. Scale variables do not have value labels since the numbers have direct meaning themselves.

The mean has some unique mathematical properties. Since every score is treated equally, it is the single point that perfectly balances all of the scores in a distribution (also known as the **point of minimized variation**).

This causes the mean to be sensitive to extreme scores as it is pulled in their direction.

Let's use a simple example to illustrate the mean's sensitivity and the concept of deviations. Suppose we have a small seminar class of nine students here at CSUF. The student's ages are: 19, 20, 20, 21, 21, 21, 22, 22, and 23. What's the mean? The sum of all of the scores is 189. Then, we divide by the number of scores, 9, to get an average age of 21.

A **deviation** is the distance between any individual score and the mean. The first student is 19. So, their deviation is calculated as: 19 - 21 = -2. That is, they are two years below the mean or two years younger than the average student. The next two deviations are -1 (20 - 21). The idea about the point of minimized variation is based on the fact that the sum of all of the deviations from the mean is always zero by definition (-2 + -1 + -1 + 0 + 0 + 0 + 1 + 1 + 2 = 0).

Now, to the idea of **sensitivity**. Suppose a tenth student joins our class. This student is the grandmother of one of the other students and is 81 years old. The sum of all of the scores becomes 270 (189 + 81). Then, we divide by the number of scores, 10, to get the new mean of 27. This one additional student inflated the mean by 6 years. The mean is being pulled in the direction of this outlier and is sensitive to extreme scores. As we work with measures of central tendency and all statistics, we need to be aware of their limitations.

As you may recall, we generated a frequency distribution for the scale variable AGE in Exhibit 2.5. Run that frequency distribution again and request the mean from the statistics dialog. Do you see that the average age of the respondents to the 2021 GSS is 52.16 years?

Earlier, we looked at educational attainment using the ordinal variable DEGREE. The GSS also asks respondents for the total number of years of formal schooling that they have completed. The EDUC variable is a scale one and ranges from 0 (no formal schooling) to 20 years of formal schooling. Exhibit 3.6 provides the mean and the frequency distribution.

Exhibit 3.6. Mean and Frequency Distribution for EDUC

Frequencies

Statistics

educ highest year of school completed

N	Valid	3966
	Missing	66
Mean		14.77

educ highest year of school completed

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 no formal schooling	9	.2	.2	.2
	1	1	.0	.0	.3
	2	2	.0	.1	.3
	3	3	.1	.1	.4
	4	1	.0	.0	.4
	5	2	.0	.1	.5
	6	15	.4	.4	.8
	7	5	.1	.1	1.0
	8	25	.6	.6	1.6
	9	32	.8	.8	2.4
	10	52	1.3	1.3	3.7
	11	83	2.1	2.1	5.8
	12	829	20.6	20.9	26.7
	13	277	6.9	7.0	33.7
	14	542	13.4	13.7	47.4
	15	208	5.2	5.2	52.6
	16	942	23.4	23.8	76.3
	17	258	6.4	6.5	82.9
	18	351	8.7	8.9	91.7
	19	113	2.8	2.8	94.6
	20	216	5.4	5.4	100.0
	Total	3966	98.4	100.0	
Missing	System	66	1.6		
Total		4032	100.0		

As we can see, the average respondent has completed 14.77 years of schooling. Notice that there are few respondents in the first half of the distribution. The cumulative percent shows that only 5.8% of respondents have completed 11 or fewer years of formal schooling. Around one-fifth (20.9%) have completed 12 years of schooling (the normative time to graduate from high school). We can also see that the median is 15 years since the 50th percentile lands in that category. The mode, the most common score, is 16 years with 23.8% of the respondents.

Earlier, in Exhibit 3.5, we compared the median of ATTEND across the five MARITAL groups and saw that there appears to be an association. When considering possible explanations, it was thought that AGE may also be a factor. So, let's split our GSS2021.sav file by MARITAL and generate the mean AGE by marital status. Again, this procedure creates a lot of SPSS output and we are only interested in the means at this point. Thus, Exhibit 3.7 provides just the statistics boxes with the average AGE by marital group ranked from low to high. Remember that you can use the navigation pane on the left side of the SPSS output viewer to quickly find exactly what you need without having to scroll through all of the other output.

Exhibit 3.7. Mean for AGE by MARITAL Groups

Statistics[a]

age age of respondent

N	Valid	908
	Missing	64
Mean		38.52

a. marital marital status = 5 never married

Statistics[a]

age age of respondent

N	Valid	87
	Missing	9
Mean		50.84

a. marital marital status = 4 separated

Statistics[a]

age age of respondent

N	Valid	1819
	Missing	180
Mean		54.38

a. marital marital status = 1 married

Statistics[a]

age age of respondent

N	Valid	608
	Missing	47
Mean		57.05

a. marital marital status = 3 divorced

Statistics[a]

age age of respondent

N	Valid	271
	Missing	30
Mean		72.18

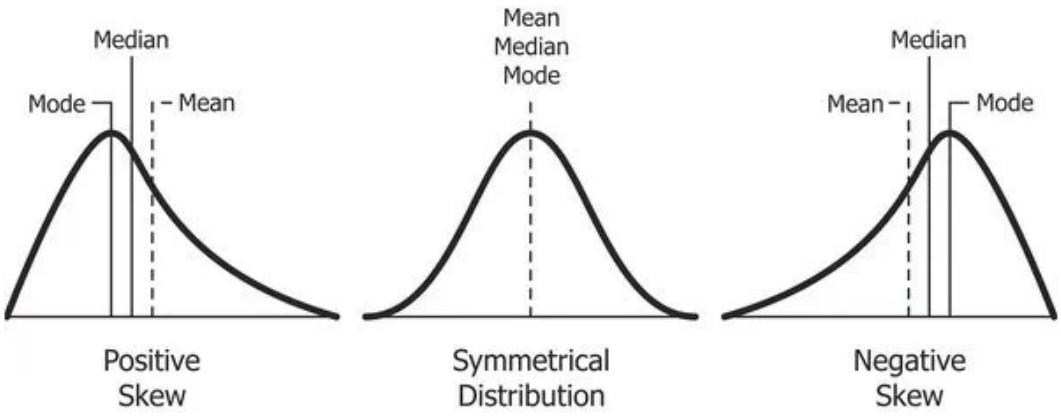
a. marital marital status = 2 widowed

The average age of respondents who have never been married is 38.52 years. Those who are separated are 50.84, married are 54.38, and divorced respondents are 57.05 years of age on average. The average age of respondents who are widowers is 72.18 years. Thus, we do see a lot of variation in the mean scores for AGE across the various MARITAL groups. Sociologists know that marriage rates have been slowly declining in the United States since the 1960s.[2] Also, the age at first marriage (for those who do get married) has been slowly rising. Thus, it makes sense that those who have never been married are the youngest. Likewise, it is sensible that widowers are the oldest on average.

Symmetrical Distributions and Skewness

The mean, median, and mode also allow survey researchers to assess the shape of the distribution of a scale variable. We produced a histogram of AGE in Exhibit 2.10 and made reference to the normal curve (to be covered in Chapter 5). A variable is normally distributed when it is bell-shaped. The normal curve is also a **symmetrical** distribution in which there is a peak in the middle ("unimodal") and the left and right sides are mirror images of one another. When the mode, the median, and the mean coincide, or are almost identical, the distribution is symmetrical (see middle image of Exhibit 3.8).

Exhibit 3.8. Symmetrical and Skewed Distributions[3]



Distributions in which the mean is lower than the median are said to be **negatively skewed**. That is, there are more cases clustered above the

mean than below it and there are some unusually low values in the distribution pulling the mean down from the median (see right image in Exhibit 3.8).

Distributions in which the mean is higher than the median are **positively skewed**. In this instance, there are more cases clustered below the mean and there are some unusually high values pulling the mean up from the median.

This chapter has provided an introduction to three measures of central tendency as descriptive statistics that assist us in describing patterns in data. We learned that the mode, the most common score in a distribution, is the appropriate measure of central tendency for nominal variables. The median, the middle case of a distribution, is used with ordinal variables. Since both nominal and ordinal variables have text-based response categories, the numeric values do not have intrinsic meaning. Therefore, when reporting the mode and median, sociologists must report both the statistic (the value) and the value label of that category (the meaning).

Mathematically, the mean is the most powerful measure of central tendency. We only use the mean with scale variables. There are no value labels or text-based categories with these since the respondents reply to such questions with numbers that have direct meaning. This chapter also employed the MARITAL variable in the split file command that we learned in Chapter 2. This allows us to start thinking about associations among variables. We also began interpreting our findings by explaining the social dynamics underlying them. The chapter concluded with a brief overview of the shapes of distributions and how the measures of central tendency are useful in assessing them. Chapter 4 is a natural extension of this one as we are now prepared to learn about measures of dispersion.

Key Terms

Mode, most, median, percentile, mean, point of minimized variation, deviation, sensitivity, symmetrical, negatively skewed, and positively skewed

Endnotes

1. Wuthnow, Robert. 2007. *After the Baby Boomers: How Twenty- and Thirty-Somethings Are Shaping the Future of American Religion*. Princeton, NJ: Princeton University Press.

2. Cherlin, Andrew J. 2010. “Demographic Trends in the United States: A Review of Research in the 2000s.” *Journal of Marriage and Family* 72 (3): 403-419.

3. Exhibit 3.8 is licensed CC BY-SA 4.0 by Diva Dugar

This 2025 work is licensed under CC BY-NC-SA 4.0

Quantitative Sociology Essentials:
An Introduction to Survey Data Analysis using IBM® SPSS®

Ed Collom, Ph.D.
Professor of Sociology
California State University, Fullerton

Chapter 4. Measures of Dispersion

Measures of dispersion complement the measures of central tendency and also assist sociologists in finding patterns in the survey data they are analyzing. They provide a single summary statistic describing the variation of scores in a distribution. Measures of dispersion quantify the concepts of heterogeneity and homogeneity that were introduced in Chapter 2. As with all of the statistics that quantitative researchers use, the level of measurement of the variable determines which measure of dispersion to employ.

Chapter Objectives

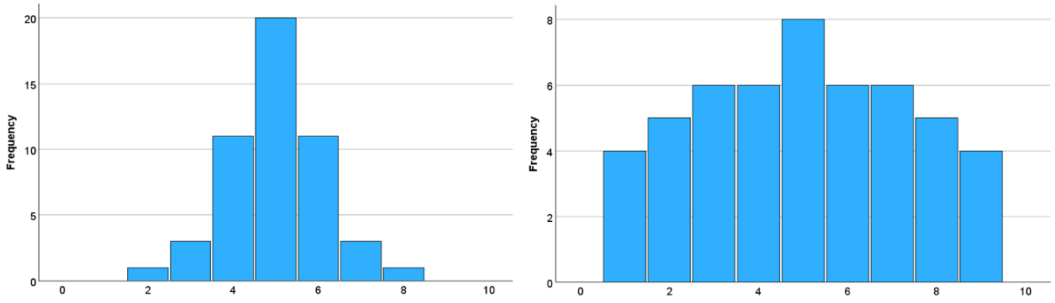
After reading this chapter, students should be able to:

- Select and produce appropriate measures of dispersion in SPSS
- Report values of measures of dispersion and describe distributions
- Compare values of measures of dispersion across groups and draw conclusions

Let's begin with an example of two different distributions to illustrate how they can vary. Suppose that 50 students in a statistics course completed two 10-point quizzes. The average performance (mean) on both quizzes was 5.0 (these were difficult quizzes!). Yet the appropriate measure of central tendency does not tell the whole story. The concept of dispersion allows us to consider the spread of the scores. Recall that homogeneity refers to sameness, the idea that there is similarity among the cases on the variable. Heterogeneity is about diversity and the extent to which there are differences among the student scores.

Exhibit 4.1 displays histograms of the 50 scores on Quiz 1 (left) and Quiz 2 (right). The values on the horizontal, X-axis are the number of correct answers out of 10 on each quiz. The frequencies of student scores are labeled on the vertical, Y-axis. We can tell by the height of the middle bar on the left histogram that 20 students scored 5 (the mean) on Quiz 1. Notice that the distribution is "peaky" as there are many scores clustered closely around the mean. This is a fairly **homogeneous distribution** as the scores are more concentrated (and note that no students scored 0, 1, 9, nor 10).

Exhibit 4.1. Histograms of Fictitious Data on Student Quiz Performance



On the other hand, we can see that student scores on Quiz 2 are much more dispersed and the distribution is flatter without so many differences in the height of the bars. Only 8 of the 50 students scored 5 on this quiz. The cases are more evenly spread out in this **heterogeneous distribution**.

Both histograms are also symmetrical. That is, the left and right-sides are mirror images of one another. In symmetrical distributions such as these, the mean, median, and mode are all identical (5). Yet if we only focused on the measures of central tendency, we would miss an important part of the statistical story. By incorporating the idea of dispersion, we can see that there are striking differences. The Quiz 1 distribution has lower dispersion and is more homogeneous. Quiz 2 has greater dispersion and is more heterogeneous.

Index of Qualitative Variation

The **index of qualitative variation** (IQV) is the appropriate measure of dispersion for nominal variables. Some of the statistics we encounter in this course have a convenient metric. The IQV is one of these. By definition, the IQV ranges from 0.00 to 1.00. A value of 0.00 indicates that all of the cases are in one category of the variable. This is the instance of no variation and maximum homogeneity (a histogram with one bar). An IQV value of 1.00 indicates that the cases are perfectly evenly distributed across all of the valid categories of the variable. This is the instance of perfect variation and maximum heterogeneity (a histogram in which all the bars are of an identical height).

While the 0.00 to 1.00 metric is convenient, the IQV for most nominal variables will fall somewhere in between and there are no strict guidelines for interpreting such values. Therefore, it is very useful to compare this statistic across multiple variables with the same metric or across multiple groups. By using the split file command in SPSS with a grouping variable, we are able to identify groups that are more homogeneous and those that are more heterogeneous on the measure of interest.

The real inconvenience and total bummer is that SPSS does not calculate the IQV! Instead of having to compute it by hand, your Canvas course website contains an Excel template that will calculate it for you. To use the IQV Calculator.xlsx file, we must first use SPSS to produce frequency distributions.

Since 1973, the GSS has included a nominal variable entitled PORNLAW:

> Which of these statements comes closest to your feelings about pornography laws? a) There should be laws against the distribution of pornography whatever the age, b) There should be laws against the distribution of pornography to persons under 18, or c) There should be no laws forbidding the distribution of pornography.

An IQV of 0.00 on this variable would indicate that all respondents fall in just one of these three categories (perfect homogeneity). An IQV of 1.00 would indicate that each of the three categories contains exactly one-third of the respondents (perfect heterogeneity). Let's split our GSS2021.sav datafile by the SEXBIRTH1 variable and see if the dispersions for males and females differ on the PORNLAW variable.

Exhibit 4.2. Frequency Distribution of PORNLAW by SEXBIRTH1

sexbirth1 r's sex assigned at birth (2021) = 1 male

pornlaw feelings about pornography laws[a]

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 there should be laws against the distribution of pornography whatever the age	191	11.0	16.9	16.9
	2 there should be laws against the distribution of pornography to persons under 18	858	49.6	76.1	93.1
	3 there should be no laws forbidding the distribution of pornography	78	4.5	6.9	100.0
	Total	1127	65.1	100.0	
Missing	System	603	34.9		
Total		1730	100.0		

a. sexbirth1 r's sex assigned at birth (2021) = 1 male

sexbirth1 r's sex assigned at birth (2021) = 2 female

pornlaw feelings about pornography laws[a]

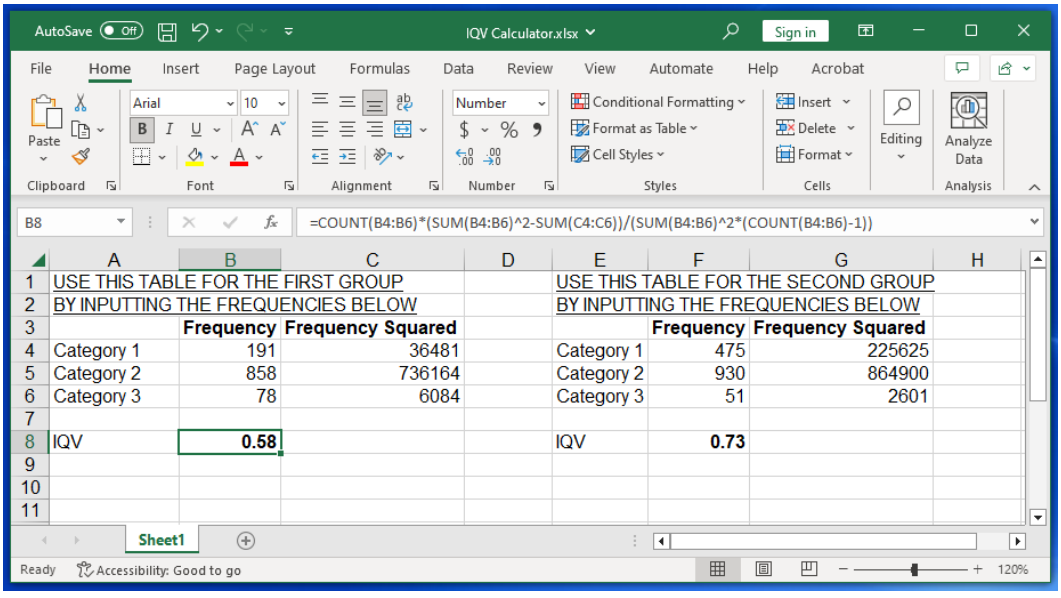
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 there should be laws against the distribution of pornography whatever the age	475	21.6	32.6	32.6
	2 there should be laws against the distribution of pornography to persons under 18	930	42.3	63.9	96.5
	3 there should be no laws forbidding the distribution of pornography	51	2.3	3.5	100.0
	Total	1456	66.2	100.0	
Missing	System	742	33.8		
Total		2198	100.0		

a. sexbirth1 r's sex assigned at birth (2021) = 2 female

Exhibit 4.2 provides the frequency distribution for males and females on PORNLAW. For males, the modal category is 2 (“There should be laws against the distribution of pornography to persons under 18”) and contains around three-quarters of the respondents. For females, the mode is also 2 with 63.9% of respondents. Yet despite the value of the appropriate measure of central tendency being the same for both groups, the frequency distributions do reveal sex differences. Males are less likely than females to have answered 1 (“There should be laws against the distribution of pornography whatever the age”) and more likely to have answered 3 (“There should be no laws forbidding the distribution of pornography”).

Let’s calculate the IQV for males and females on the PORNLAW variable to quantify the differences in these distributions. The IQV Calculator.xlsx file is a template for easy calculation of the IQV. It contains five response categories for the nominal variable of interest. In this case, the variable only has three valid categories. Therefore, we can simply delete rows 7 (Category 4) and 8 (Category 5) of the spreadsheet.

Exhibit 4.3. Calculation of IQV for Males and Females on PORNLAW



	A	B	C	D	E	F	G	H
1	USE THIS TABLE FOR THE FIRST GROUP				USE THIS TABLE FOR THE SECOND GROUP			
2	BY INPUTTING THE FREQUENCIES BELOW				BY INPUTTING THE FREQUENCIES BELOW			
3		Frequency	Frequency Squared			Frequency	Frequency Squared	
4	Category 1	191	36481		Category 1	475	225625	
5	Category 2	858	736164		Category 2	930	864900	
6	Category 3	78	6084		Category 3	51	2601	
7								
8	IQV	0.58			IQV	0.73		
9								
10								
11								

To use this calculator, we need only enter the raw frequencies of the three valid categories into the “Frequency” column for the first group (males)

and the second group (females). Using the data from Exhibit 4.2, Exhibit 4.3 illustrates the completed spreadsheet for this particular IQV calculation.

The IQV for males on PORNLAW is 0.58 and for females it is 0.73. Having group IQV values to compare to one another makes it much easier to interpret the extremes between 0 (perfect homogeneity) and 1 (perfect heterogeneity). Here, we see that females have a more heterogeneous distribution on this variable than males. As noted earlier, there are fewer females in category 2 and more females in category 1. This makes their dispersion greater than that of males.

Range and Interquartile Range

There are two appropriate measures of dispersion for ordinal variables. The range and the interquartile range complement each other and are frequently used together. The **range** is the distance between the highest and the lowest scores in a distribution. It is easy to calculate: high score - low score = range. As a statistic, the range has limitations since it only focuses upon the extreme scores.

The **interquartile range** should be used with the range to gain more information about dispersion on an ordinal variable. The interquartile range is the range of values comprising the middle 50% of a distribution. Its reference point is the median, the exact midpoint or 50th percentile of a distribution (recall that the median is the appropriate measure of central tendency for ordinal variables). As indicated in its name, quartiles are the basis of the interquartile range.

Quartiles are the value at or below which one-quarter of the respondents fall. They are the result of dividing up a distribution into four equal parts (quarters). Quartiles are based upon the same logic as the cumulative percent in frequency distributions. Quartile 1 (also known as the 25th percentile) is the value of the lower quartile of a distribution in which the first quarter of the respondents are located. Quartile 2 is the value of the 50th percentile, the median. This second quarter of the distribution encompasses the cases from the 25th percentile up to the 50th percentile. Quartile 3 (also known as the 75th percentile) is the value at which the third quarter of the distribution ends. Quartile 4 is the highest score in a distribution, representing the upper-limit of the last quarter of the distribution which began at the 75th percentile.

Once the values of the quartiles are calculated, the interquartile range is easily computed: Quartile 3 - Quartile 1 = interquartile range. This difference between the lower and upper quartiles gives us the range of values comprising the middle 50% of a distribution. By definition, the median is the midpoint of the interquartile range.

Exhibit 4.4 revisits the POLVIEWS variable introduced in Chapter 2. As we saw in Chapter 3, the Analyze > Descriptive Statistics > Frequencies command in SPSS contains a Statistics dialog enabling us to request measure of central tendency and measures of dispersion (but not the IQV!). In that dialog, the Range is found in the bottom-left corner under Dispersion. To compute the interquartile range, we must also request Quartiles from the top-left corner under Percentile Values.

Exhibit 4.4. Range and Interquartile Range for POLVIEWS

Frequencies

Statistics

polviews think of self as liberal or conservative

N	Valid	3964
	Missing	68
Range		6
Percentiles	25	3.00
	50	4.00
	75	5.00

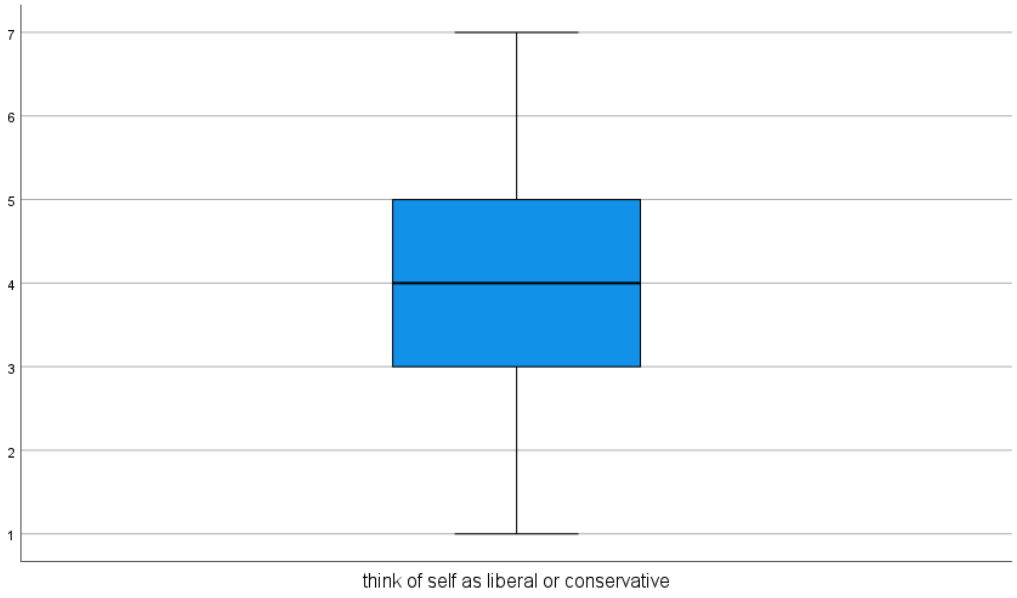
polviews think of self as liberal or conservative

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 extremely liberal	207	5.1	5.2	5.2
	2 liberal	623	15.5	15.7	20.9
	3 slightly liberal	490	12.2	12.4	33.3
	4 moderate, middle of the road	1377	34.2	34.7	68.0
	5 slightly conservative	476	11.8	12.0	80.0
	6 conservative	617	15.3	15.6	95.6
	7 extremely conservative	174	4.3	4.4	100.0
	Total	3964	98.3	100.0	
Missing	System	68	1.7		
Total		4032	100.0		

The Statistics box of the exhibit reports the range as 6. The range is simply the distance between the highest (7, “extremely conservative”) and the lowest (1, “extremely liberal”) scores in a distribution (7 - 1 = 6). The value of Quartile 3 is listed as the 75th percentile: 5.00. Notice in the frequency distribution that the Cumulative Percent indicates that category 5 (“slightly conservative”) does contain the 75th percentile. The value of Quartile 1 is listed as the 25th percentile: 3.00. Again, notice that the Cumulative Percent shows that category 3 (“slightly liberal”) does contain the 25th percentile. And, the median = 4 (“moderate, middle of the road”) since that is the category containing the 50th percentile. Thus, the interquartile range is 5 - 3 = 2.

So, we use the Statistics box to compute the interquartile range of POLVIEWS as 2 and we see that the range is reported as 6. What does this mean as far as quantifying dispersion? When the interquartile range is fairly small in relation to the range, it indicates a more homogeneous (less dispersed) distribution. The interquartile range of 2 means that half or more of all of the respondents chose either category 3, 4, or 5. In other words, the small interquartile range does show some clustering of cases near the median and suggests a fairly homogeneous distribution.

Exhibit 4.5. Boxplot of POLVIEWS



We can also visualize the range and interquartile range using **boxplots**. The command is Graphs > Boxplot… The Boxplot dialog has two options. At the top, we want a Simple boxplot and at the bottom we want "Summaries of separate variables." Once you select those options, click Define. Now, simply move the POLVIEWS variable into the Boxes Represent box on top and click OK. Exhibit 4.5 displays the result.

The blue box in the boxplot is the interquartile range, the middle 50% of the distribution (notice the value of 3 on the Y-axis for Quartile 1 and the value of 5 for Quartile 3). The line in the middle of the box is the median (4). These are useful graphs as we can visualize the dispersion and know that a small box relative to the size of the graph (the range) suggests a more homogeneous distribution.

Let's explore some other ordinal variables and their dispersion. The 2021 version of the General Social Survey introduced a new series of questions about trust in institutions:

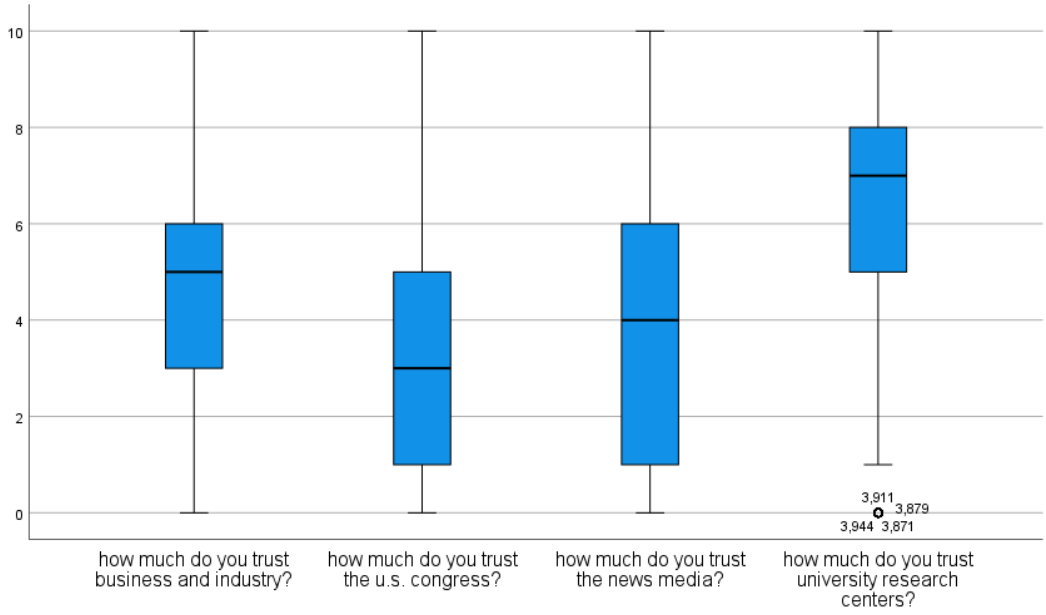
> On a scale of 0 to 10, how much do you personally trust each of the following institutions? 0 means you do not trust an institution at all, and 10 means you trust it completely. Business and industry (TRBUSIND); The U.S. Congress (TRLEGIS); The news media (TRMEDIA); University research centers (TRRESRCH)

Do not get confused by the format of the response categories, "on a scale of 0 to 10." These are ordinal variables, not scale ones. Remember that the latter are numbers with direct meaning in their original metric. The values within this 0 to 10 trust scale cannot be treated as numbers with meaningful units. That is, someone who answered 6 is higher and has more trust in that institution than someone who answered 5, but that difference of 1 is not a mathematically meaningful increase of one unit. Whenever the prompt of the survey question equates text with the numbers ("0 means you do not trust…"), you will know that the variable is an ordinal one.

Exhibit 4.6 displays boxplots for the four variables on one graph. To produce this on your own, move all four variables into the Boxes Represent box at the top of the dialog. This is a useful way for us to make comparisons on these variables with identical response categories. We can see that all four have a range of 10 as some respondents answered in the extreme categories of 0 and 10 on each. Those who chose 0 on

TRRESRCH look different on the graph since they are considered outliers given the high value of the median.

Exhibit 4.6. Boxplots of TRBUSIND, TRLEGIS, TRMEDIA, & TRRESRCH



Using the values of the Y-axis on the graph to identify the size of the boxes, we can see that the interquartile range is 3 for TRBUSIND, 4 for TRLEGIS, 5 for TRMEDIA, and 3 for TRRESRCH. So, the distributions on TRBUSIND and TRRESRCH are more homogeneous while the one for TRMEDIA is more heterogeneous. In looking at the median values, we see that respondents have the highest level of trust in university research centers and the lowest level of trust in the U.S. Congress.

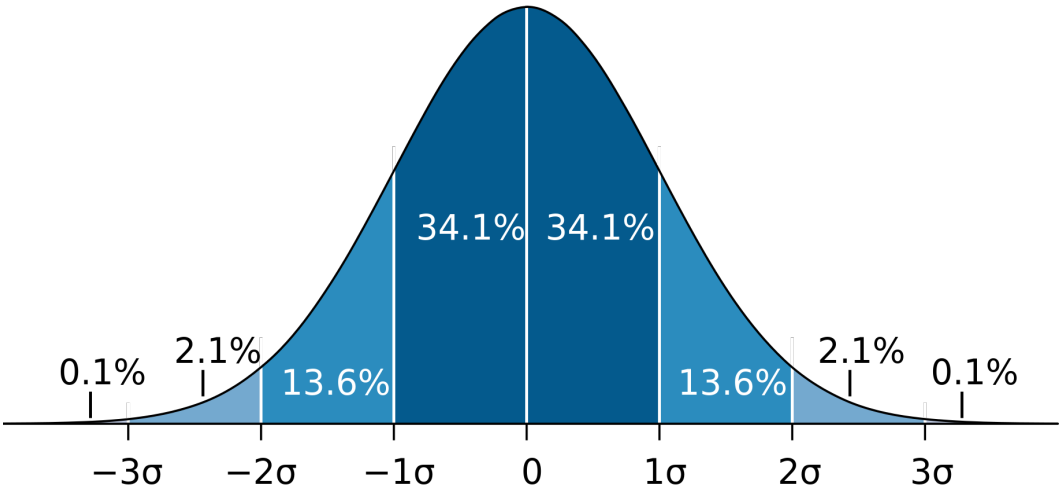
Standard Deviation

The **standard deviation** is the most widely used measure of dispersion and the appropriate one for scale variables. Its computation is based upon deviations, the distance between each individual score and the mean (as discussed in Chapter 3). This statistic does not have a fixed metric. While zero is its lowest possible value (indicating perfect homogeneity), standard deviation scores can be very large values since it is expressed in the units of the variable.

As with the other measures of dispersion, it is useful to compare standard deviation values across different groups. Smaller standard deviations indicate more homogeneous distributions while larger ones represent more heterogeneous distributions. The range also provides a convenient comparison point. We may have distributions for two groups with similar standard deviation values, but different ranges. The group with the larger range would have the more homogeneous distribution since the cases in that distribution would be more clustered around the mean. This logic is very similar to having a small interquartile range compared to the range with ordinal variables.

As foreshadowed previously, the next chapter will cover the concept of the normal curve. The bell-shaped, symmetrical normal distribution can be used with the standard deviation to make precise measurements of the dispersion of responses in relation to the mean. Exhibit 4.7 illustrates the **area under the normal curve** in standard deviation units. The mean (the peak of the normal curve) is 0 when working with standard deviation units. In a normal distribution, 68.26% of the cases fall within one standard deviation unit of the mean. That is, 34.13% of the area under the normal curve always falls within one standard deviation above the mean and another 34.13% falls within one standard deviation unit below the mean. Cases become increasingly rare beyond that as 95.44% fall within two standard deviation units of the mean.

Exhibit 4.7. Area under the Normal Curve in Standard Deviation Units[1]



Let's calculate our first standard deviation using SPSS. As you may have noticed, the standard deviation is available from the Statistics dialog of the Analyze > Descriptive Statistics > Frequencies command. There is another way to quickly get basic descriptive statistics from any variable. The Analyze > Descriptive Statistics > Descriptives command is useful if you do not need a frequency distribution. Go to this command, move AGE into the Variable(s) box, and click OK.

Exhibit 4.8 provides the output. We see that the average respondent to the 2021 GSS is 52.16 years of age. The value of the standard deviation is 17.23. The Minimum and Maximum values on the variable are also provided, permitting easy calculation of the range. If we assume that AGE is normally distributed, we can add and subtract the standard deviation to and from the mean to calculate the areas under the normal curve. 52.16 + 17.23 = 69.39. So, assuming normality, 34.13% of the cases fall between 52.16 and 69.39 years of age. Another 34.13% of respondents are within one standard deviation unit below the mean: between 34.93 and 52.16 years of age. Altogether, 68.26% of the cases would be between 34.93 and 69.39 years of age.

Exhibit 4.8. Analyze > Descriptive Statistics > Descriptives Command Output for AGE

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
age age of respondent	3699	18	89	52.16	17.233
Valid N (listwise)	3699				

Now, let's use the Split File command again to determine if marital status groups vary in the amount of television that they watch. The TVHOURS variable asks respondents to report the number of hours that they watch television in a typical day. Split the GSS2021.sav datafile by MARITAL and then run the Analyze > Descriptive Statistics > Descriptives command for TVHOURS.

Exhibit 4.9. Descriptives Output for TVHOURS by MARITAL

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
tvhours hours per day watching tv	1338	0	24	3.20	2.756
Valid N (listwise)	1338				

a. marital marital status = 1 married

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
tvhours hours per day watching tv	197	0	24	4.43	3.387
Valid N (listwise)	197				

a. marital marital status = 2 widowed

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
tvhours hours per day watching tv	452	0	24	3.88	3.384
Valid N (listwise)	452				

a. marital marital status = 3 divorced

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
tvhours hours per day watching tv	65	0	24	4.34	4.731
Valid N (listwise)	65				

a. marital marital status = 4 separated

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
tvhours hours per day watching tv	626	0	24	3.31	3.222
Valid N (listwise)	626				

a. marital marital status = 5 never married

Exhibit 4.9 provides the output. First, we can see that the range for all five groups is 24. Some people do not watch any television (0), while at least one respondent from each of the five marital groups responded with 24 hours. Some people do have the television on 24/7, but we all have to sleep some time! Since the ranges are all the same, we can simply compare the values of the standard deviations. Notice that married respondents have the lowest standard deviation at 2.76. This indicates that they have the most homogeneous distribution (they also report watching the least amount of TV with a mean of 3.20 hours). Respondents who are separated have the largest standard deviation (4.73) and the most heterogeneous distribution. In other words, separated people vary a lot in how much television they watch.

In this chapter we learned that measures of dispersion are statistics quantifying the variation within a distribution. Homogeneity and heterogeneity are key terms allowing us to describe distributions. We choose our measures of dispersion based upon the variable's level of measurement. The index of qualitative variation ranges from 0.00 to 1.00 and is to be used with nominal variables. Unfortunately, it is not available in SPSS, so we will have to use the Excel calculator. The range and interquartile range measure the distance between the high and low values of a variable and the distance comprising the middle 50% of the cases. These measures of dispersion are to be used with ordinal variables. The standard deviation is the appropriate measure of dispersion for scale variables. It is the most mathematically powerful one given the properties of the normal curve. In the next chapter, we will see how the normal distribution is the basis from which we are able to make inferences about our data.

Key Terms

Homogeneous distribution, heterogeneous distribution, index of qualitative variation (IQV), range, interquartile range, quartiles, boxplots, standard deviation, area under the normal curve

Endnotes

1. Exhibit 4.7 is licensed CC BY-SA 4.0 by M.W. Toews

This 2025 work is licensed under CC BY-NC-SA 4.0

Quantitative Sociology Essentials: An Introduction to Survey Data Analysis using IBM® SPSS®

Ed Collom, Ph.D.
Professor of Sociology
California State University, Fullerton

Chapter 5. Making Inferences

Thus far we have focused on descriptive statistics from the 2021 General Social Survey. In this chapter we will learn about inferential statistics and making generalizations from sample data to the larger population being represented. We will begin with a discussion of sampling, return to the idea of the normal distribution, introduce Z-scores, and conclude with the concept of the sampling distribution.

Chapter Objectives

After reading this chapter, students should be able to:
- Explain the difference between populations and samples
- Express familiarity with sampling methods, bias, and error
- Compute and report Z-scores
- Understand the purpose of inferential statistics and the sampling distribution

Populations, Samples, and Sampling

A **population** is the total set of elements that the sociologist is interested in studying. An **element** is an entity of the population and is most frequently an individual when conducting quantitative sociological research. While some do investigate groups, events, or objects, sociologists usually study individual people. Though we are often interested in entire populations, they are typically impossible to study. Many populations are comprised of too many individuals, making it impractical to contact all of them. In some cases, we do not even have a list of the elements nor their contact information. We also know that some people will not want to participate in our study. Even if we had a list of the

population with their contact information, some would not take our survey, resulting in incomplete population data.

To overcome the difficulties of studying populations, survey researchers study samples instead. A **sample** is a carefully selected subset of cases from a population. The GSS is nationally representative sample of noninstitutionalized, English or Spanish-speaking American adults. The process of identifying and selecting the subset of the population for study is known as **sampling**. The ultimate goal of sampling is to have a random subset that closely resembles the characteristics of the overall population.

Sociologists collect and analyze data from a sample in order to make generalizations to the larger population being represented. **Inferential statistics** involves making predictions about a population using observations from a sample. This body of statistics involves quantifying the likelihood of being wrong, or committing an error, when making generalizations or failing to make them when they should be made.

In order to engage in inferential statistics and make generalizations, sociologists must use **probability sampling** methods to systematically select the cases comprising their sample. This scientific approach enables calculation of the probability of each case in the population being included in the sample. Survey researchers studying American adults usually aim for a minimum of 1,000 to 1,500 respondents to complete the survey.[1] Using sound sampling methods, this **sample size** is adequate to represent the 264 million residents of the United States aged 18 or older that typically comprise the population of interest for nationally representative surveys.[2]

The **response rate** is the percent of the individuals in the selected sample that actually complete the survey. Since some people who are invited do not respond, survey researchers must anticipate nonresponse and invite extra individuals in their **selected sample** to reach their minimum sample size. After the survey has closed, the sampling experts compare the characteristics of the **respondents** with the individuals comprising the selected sample. Since the selected sample was specifically chosen to represent American adults, **nonresponse bias** occurs when respondents are different than those who did not respond. This bias is one form of **sampling bias** which occurs when the characteristics of the sample do not match those of the larger population. Sampling bias can be the result

of errors in the sampling process that end up excluding potential participants.

One potential solution to nonresponse and sampling bias is widely used in survey research. Sampling experts usually create **weights** in survey datasets to correct for these biases. This topic goes beyond the content of this course and we will not employ any weights in our analyses. However, the idea is that the size of under- and over-represented groups in the sample is adjusted to match population totals from trusted agencies such as the U.S. Census Bureau. As described in the codebook, the 2021 GSS includes a weighting variable (WTSSNRPS) that adjusts the sample for nonresponse bias and to match the respondent totals on education, sex, age, region, race, and ethnicity to that of the U.S. population. However, weights are typically not used in hypothesis testing of associations among variables (the core topic of the majority of this text).

Sampling error is another aspect of survey research. It is inherent with the use of samples since they each vary from one another. Sampling error is quantified to estimate the extent to which characteristics of the sample do not match those of the larger population. As is likely evident by now, sampling is a science in and of itself and will not be covered deeply here. For now, let's briefly consider the five major probability sampling methods.

The **simple random sample** is one in which every member of the population and every possible combination of members has an equal chance of being selected. Drawing names out of a hat is the classic example. With a numbered list of the individuals comprising the population, we could also use a random number generator to randomly choose the number of cases that we desire in our selected sample.

Systematic random sampling is similar to simple random sampling, except that only the first case is randomly drawn. Here, the sociologist calculates an interval (K) from which to select cases. Suppose the list of our population elements contains 1,000 people and that we have determined that the size of our selected sample should be 100. One-thousand divided by 100 results in an interval of 10. After starting randomly in the list, we would then select every 10^{th} case to reach our selected sample size of 100. It is important that the list of the population elements (known as the **sampling frame**) is random and is not systematically arranged by any characteristic of the individuals.

Proportionate stratified sampling is when we divide the population into multiple sampling frames or lists based on key variables of our study. When we know the population percentages on these variables, we can match our selected sample to these proportions. Suppose we are interested in studying CSUF student attitudes about the racial climate on campus. We would want to assure that the racial composition of our selected sample reflects the racial composition of the student body. The Institutional Research website allows anyone to query such data. In Spring 2025, 22,360 of the 41,040 enrolled students identified as "Hispanic/Latino." This is 54.5% of the student body. Thus, if we chose 1,000 as our selected sample size, we would invite 545 "Hispanic/Latino" students to complete our survey under the proportionate stratified approach.

Disproportionate stratified sampling uses the same logic, but selects more cases from groups that are a small proportion of the population. In Spring 2025 there were only 1,038 enrolled students identifying as "Black or African American." This is 2.5% of the student body. The proportionate stratified sample approach to a selected sample size of 1,000 would lead us to invite only 25 "Black or African American" CSUF students to take our survey. As we shall see at the end of this chapter, inferential statistical analyses require more cases than this. Thus, under the disproportionate stratified approach, we may decide to invite 100 "Black or African American" students to participate to better learn of their campus experience and enable inferential statistical analyses.

As you may have noticed, these first four sampling methods all require a sampling frame. We are not always fortunate enough to have a list of the individuals comprising our population. This is the case for nationally representative surveys as no accurate list of U.S. residents exists. **Cluster sampling** is based on geography and the selection of household addresses within the United States. There are databases, including that of the United States Postal Service, that experts fielding surveys such as the GSS use to randomly select household addresses that are representative of the various geographic regions in the U.S. Thus, the unit of analysis for cluster sampling purposes is the residence. Then, a specific individual within that residence is invited to participate. Notice that this approach does omit those who are unhoused or living within an institution such as a prison.

Normal Distribution and Z Scores

The normal distribution is the perfectly bell-shaped, symmetrical distribution referenced earlier in the text. Here, the mean, median, and mode are identical and coincide with the value at the peak of the distribution. The normal curve is a theoretical distribution and tool that survey researchers employ. There is no empirical variable that has a perfect normal distribution.

In the previous chapter, Exhibit 4.7 illustrates the normal distribution in standard deviation units. By definition, the standard normal distribution has a mean of 0.0 and a standard deviation of 1.0. **Z scores** are variables that have been transformed to the normal distribution. They are referred to as standardized scores since the metric of the original variable has been aligned with the normal curve. After transformation, variables expressed as Z scores are in standard deviation units and represent how far a raw score is from the mean. Positive Z scores reflect cases that are above the mean and negative Z scores are those below the mean.

Consider the three scale variables that we have already investigated from the 2021 GSS: AGE, EDUC, and TVHOURS. Running the Analyze > Descriptive Statistics > Descriptives command with these three variables produces the output in Exhibit 5.1. We see that the average respondent is 52.16 years of age, has 14.77 years of formal education, and watches 3.46 hours of television per day.

Exhibit 5.1. Descriptives Output for AGE, EDUC, and TVHOURS

Descriptives

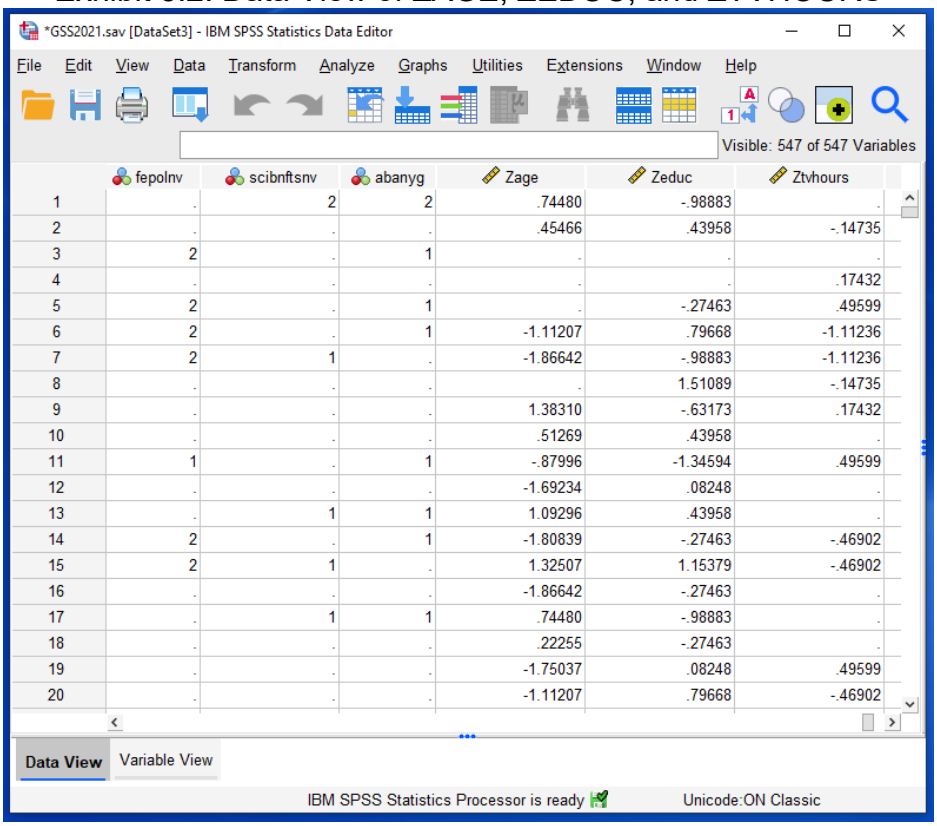
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
age age of respondent	3699	18	89	52.16	17.233
educ highest year of school completed	3966	0	20	14.77	2.800
tvhours hours per day watching tv	2683	0	24	3.46	3.109
Valid N (listwise)	2463				

This same dialog in SPSS is where we generate Z scores for variables. In the bottom left hand corner there is a checkbox “Save standardized values

as variables." If you select that and then hit OK, three new variables will appear in the Variable View: ZAGE, ZEDUC, and ZTVHOURS. These are the new variables that are the result of standardizing the original ones to the normal curve. Now, go back to Analyze > Descriptive Statistics > Descriptives and hit Reset. Move ZAGE, ZEDUC, and ZTVHOURS into the Variable(s) box and hit OK. Notice that the mean for all three standardized variables is 0.0 and the standard deviation is 1.0.

Exhibit 5.2. Data View of ZAGE, ZEDUC, and ZTVHOURS



	fepolnv	scibnftsnv	abanyg	Zage	Zeduc	Ztvhours
1	.	2	2	.74480	-.98883	.
2	.	.	.	.45466	.43958	-.14735
3	2	.	1	.	.	.
4	.	.	.	.	.	.17432
5	2	.	1	.	-.27463	.49599
6	2	.	1	-1.11207	.79668	-1.11236
7	2	1	.	-1.86642	-.98883	-1.11236
8	.	.	.	.	1.51089	-.14735
9	.	.	.	1.38310	-.63173	.17432
10	.	.	.	.51269	.43958	.
11	1	.	1	-.87996	-1.34594	.49599
12	.	.	.	-1.69234	.08248	.
13	.	1	1	1.09296	.43958	.
14	2	.	1	-1.80839	-.27463	-.46902
15	2	1	.	1.32507	1.15379	-.46902
16	.	.	.	-1.86642	-.27463	.
17	.	1	1	.74480	-.98883	.
18	.	.	.	.22255	-.27463	.
19	.	.	.	-1.75037	.08248	.49599
20	.	.	.	-1.11207	.79668	-.46902

Exhibit 5.2 displays the Data View screen with the Z score equivalents of AGE, EDUC, and TVHOURS for the first twenty respondents. The respondent in the first row has a score of .74480 on ZAGE. As seen above, the original AGE variable has a mean of 52.16 and a standard deviation of 17.233. By adding the standard deviation to the mean, we get 69.363. This would be the age for a respondent that is one standard deviation unit higher than the mean. The Z score of .74480 indicates that

the first respondent is above the mean, but not quite one standard deviation above. If you scroll back to the very beginning of the variables in the Data View, you can see that this first respondent is 65 years of age. Z scores change the metric of the original variable and standardize it to the normal curve.

That same first respondent has a score of -.98883 on ZEDUC. This person is almost exactly one standard deviation unit below the mean on the EDUC variable. As seen in Exhibit 5.1, EDUC has a mean of 14.77 and a standard deviation of 2.8. Subtracting the standard deviation from the mean results in a value of 11.97 years of schooling. Find EDUC back towards the beginning of the Data View and you will see that this first respondent reported achieving 12 years of formal schooling.

Z scores are useful as they quickly show us where any one respondent falls in relation to all others. Notice that the case in the second row in Exhibit 5.2 is someone who is older than the average, more educated than the average, and one who watches slightly less television on average. The sixth respondent is younger, more educated, and watches less TV than the average. Finally, look at the twelfth case on ZEDUC. The Z score of .08248 is very close to zero, the mean. Scanning back to the original EDUC variable, we see that this respondent reported 15 years of formal education (just slightly above the average of 14.77).

Parameters and Statistics

In the next chapter we will see how the logic of the normal curve and Z scores are applied to estimating population parameters. As discussed at the beginning of this chapter, populations are typically impossible to study in survey research. We use sampling to carefully select members that represent a larger population. Survey researchers are interested in the characteristics of those in the sample because they represent a population. We really want to know about the **population parameters**. A parameter is a numerical measure of a population distribution. The problem is that we can usually only estimate parameters since we lack complete population data. **Sample statistics** are the numerical measures from our sample distribution.

Inferential statistics is the process of estimating unknown population parameters using sample statistics. The sample statistics within the 2021 GSS datafile that we have been analyzing are interesting to us only

insofar as they represent the larger population of American adults. As seen in Exhibit 5.3, we use different **symbols** to reference sample statistics and population parameters.

Exhibit 5.3. Symbols for Statistics and Parameters

	Sample Statistic	Population Parameter
Mean	$\bar{X}$	μ
Standard Deviation	s	σ
Proportion	p	π

We learned about the mean as the appropriate measure of central tendency for scale variables in Chapter 3. When calculating it in SPSS, we were producing the sample mean, $\bar{X}$ ("X bar"). However, we are really interested in the population mean, μ ("mew"). In Chapter 4, we learned that the standard deviation is the appropriate measure of dispersion for scale variables. The sample standard deviation is simply referred to as "s." The population standard deviation is known as σ ("sigma"). Finally, we've already been working with percentages in frequency distributions. In the next chapter, we will learn about estimating π ("pie"), a population proportion, using a sample proportion, p.

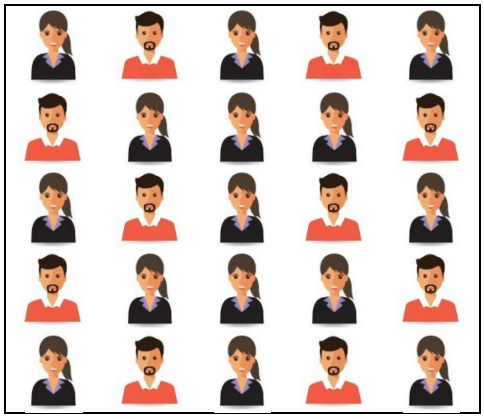
As discussed earlier, sampling error occurs when the characteristics of the sample do not match those of the larger population. In other words, sampling error is the discrepancy between a sample estimate of a population parameter and the real population parameter. We seek to minimize the error and increase the confidence that we have to infer from the sample to the population.

Exhibit 5.4 provides an example of sampling error concerning the sex composition of a classroom population. Suppose that we have a classroom of 25 students. Fifteen out of the 25 students are female. Thus, the population proportion π is .60 (15/25). If we drew a random sample of six students from this population of 25, it would not be possible to perfectly reflect the sex composition of the population. If four of the students selected for our sample were female, the sample proportion, p,

would be .67 (4/6). The difference between this sample statistic of .67 and the population parameter of .60 is the sampling error. In all of the inferential statistics that we produce, we will be accounting for the potential of sampling error.

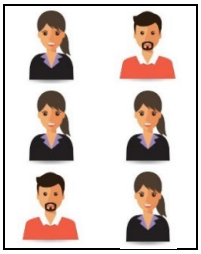
Exhibit 5.4. Sampling Error: Sex Composition of Sample differs from Classroom Population

Population



Proportion of Female Students in Population
$\pi = 15/25 = .60$

Sample



Proportion of Female Students in Sample
$p = 4/6 = .67$

Sampling Distribution

The earlier content of this chapter has all led to this final topic. The concept of the sampling distribution is the foundation of inferential statistics in survey research. The **sampling distribution** is a theoretical probability distribution of all possible sample values for the statistics of interest. That is, imagine that we repeatedly drew samples of the same population. If we did this over and over, until all possible random samples of size N are drawn from a population with a mean μ and a standard deviation σ, we could plot the sampling distribution of these sample means.

The Central Limit Theorem is the mathematical basis of the sampling distribution. It states that as the sample size N becomes larger, the

sampling distribution of the sample means becomes approximately normal in shape. This is not pertaining to the shape of the distribution of some scale variable in a population. Rather, it is concerning the probability that the sampling distribution of repeatedly drawn samples of a sufficient size will become a normal curve. Then, the mean of that sampling distribution will be equivalent to the population mean. The rule of thumb that sociologists use in determining what counts as a "large" sample is 50. That is, with a sample size of 50 or greater, the Central Limit Theorem applies and we can calculate the confidence we have in generalizing from such a sample to the larger population being represented.

This chapter has been somewhat technical and theoretical. The goal has been to introduce students to the foundations of inferential statistics without overwhelming you with formulas and abstract examples. Quantitative sociologists engaged in survey research always need to keep in the back of their minds the fact that we are analyzing samples when we are really interested in the populations that are being represented. The concepts of sampling, sampling error, and estimation of population parameters will be applied from here on out. In the next chapter, we will apply our knowledge of inferential statistics and learn about point estimates and confidence intervals.

Key Terms

Population, element, sample, sampling, inferential statistics, probability sampling, sample size, response rate, selected sample, respondents, nonresponse bias, sampling bias, weights, sampling error, simple random sample, systematic random sampling, sampling frame, proportionate stratified sampling, disproportionate stratified sampling, cluster sampling, Z scores, population parameters, sample statistics, symbols, and sampling distribution

Endnotes

1. Hibberts, Mary, R. Burke Johnson, and Kenneth Hudson. 2012. "Common Survey Sampling Techniques." In Lior Gideon (ed.) *Handbook of Survey Methodology for the Social Sciences*, pp. 53-74. New York, NY: Springer.

2. United States Census Bureau. 2025. “National Population by Characteristics: 2020-2024.” https://www.census.gov/data/tables/time-series/demo/popest/2020s-national-detail.html

This 2025 work is licensed under CC BY-NC-SA 4.0

Quantitative Sociology Essentials: An Introduction to Survey Data Analysis using IBM® SPSS®

Ed Collom, Ph.D.
Professor of Sociology
California State University, Fullerton

Chapter 6. Estimation

In this chapter, we will learn about making inferences for single variables. Throughout the text so far, we have been engaged in the process of estimation as we have analyzed frequency distributions, measures of central tendency, and measures of dispersion. A **point estimate** is a characteristic of a sample that is used to estimate or represent unknown population parameters. Thus, we have been producing a variety of point estimates in SPSS already. The problem with point estimates is that their accuracy is unclear since there is always some sampling error when we analyze sample data such as the 2021 General Social Survey. A solution to this dilemma is found with confidence intervals.

Chapter Objectives

After reading this chapter, students should be able to:
- Explain the limitations of point estimates from samples
- Construct confidence intervals and report their results
- Prepare nominal and ordinal variables for confidence interval analyses
- Understand the concept of statistical differences when comparing groups

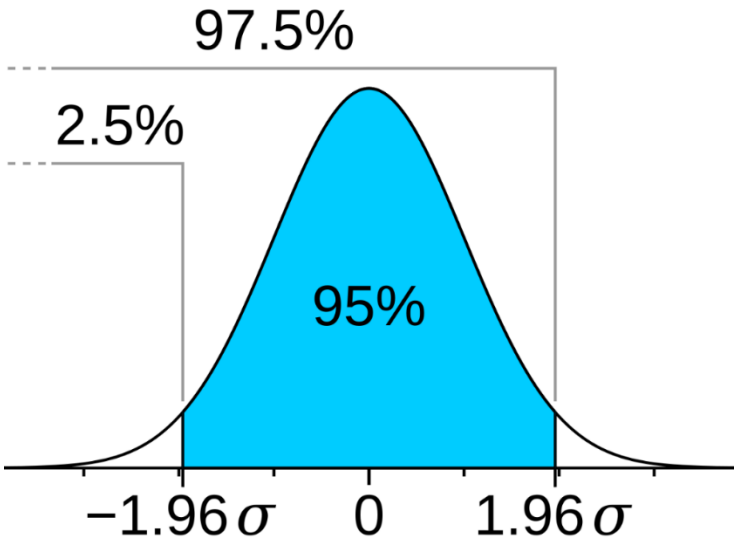
Confidence Intervals

Confidence intervals specify a range of values surrounding the point estimate within which the population parameter is estimated to fall. We begin by setting the **confidence level**, the likelihood that a specified interval will contain the population parameter. In sociology and in SPSS, the default confidence level is 95%. With a 95% confidence level, we have a margin of error of 5%. When working with sample data, there is always a **margin of error**, the risk that we are willing to take of being

wrong. That is, the risk of our confidence interval surrounding the point estimate from our sample not containing the true population parameter.

When we construct a 95% confidence interval, the true population parameter will be included within that interval 95 times out of 100. Every five times out of 100 our confidence interval will be inaccurate and will not include the population parameter. Given the central limit theorem and the characteristics of the normal distribution, we know that 95% of all random sample means will fall within ±1.96 standard deviation units of the true population mean. As seen in Exhibit 6.1, the 5% margin of error is split between the two tails of the sampling distribution. In rare cases, our sample means will be much smaller or much larger than the population mean and will not be accurate estimates of it.

Exhibit 6.1. Margin of Error in the Tails of the Sampling Distribution[1]



The idea of 95% of cases falling within ±1.96 standard deviation units of the mean takes us back to the concept of Z scores that we learned in Chapter 5. Recall that Z scores are variables standardized to the normal curve with a mean of 0.0 and a standard deviation of 1.0. The Z value is a component of the formula to calculate confidence intervals. Since we will always be working with 95% confidence intervals here, our Z will always be set at 1.96. We multiply that 1.96 by what is known as the **standard**

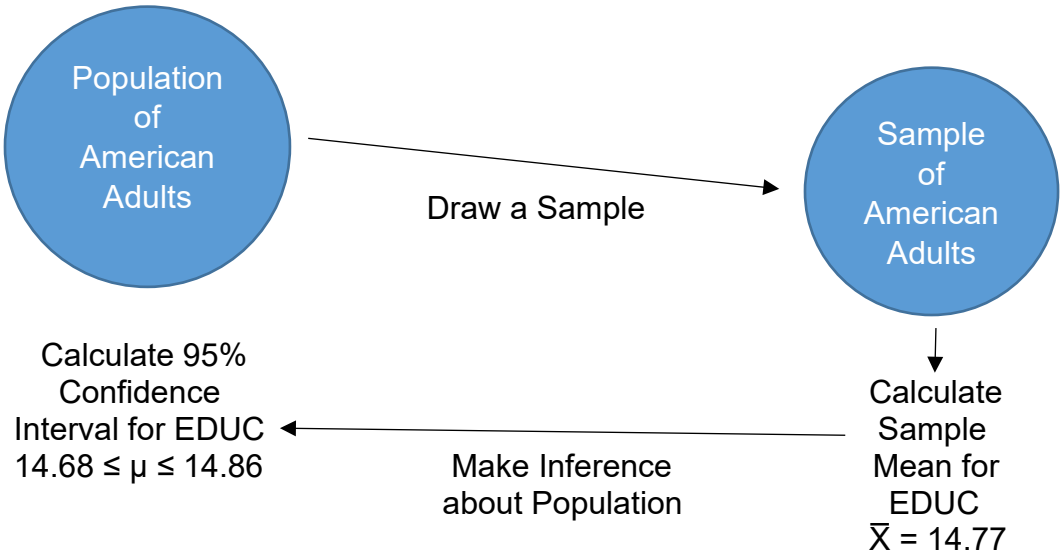
error of the mean: the sample standard deviation divided by the square root of the sample size. Once we calculate the product of Z and the standard error, we simply add it to and subtract it from the sample mean to get the range of values comprising the confidence interval. While we will not be calculating confidence intervals by hand, Exhibit 6.2 provides the formula so that students understand what SPSS is doing for us.

Exhibit 6.2. Formula for Calculating 95% Confidence Interval

$$CI = \bar{X} \pm 1.96 \left({}^{s}/_{\sqrt{N}} \right)$$

Let's recap what we've covered thus far. Exhibit 6.3 summarizes the process of constructing 95% confidence intervals. First, a sample, such as the 2021 GSS, is drawn from the population of American adults. Then, we calculate a point estimate of a variable of interest.

Exhibit 6.3. Constructing Confidence Intervals from Sample Data



As we have previously seen, the sample mean ($\bar{X}$) for the years of formal schooling variable (EDUC) is 14.77 years. To make an inference about

the true population mean (μ), we calculate a 95% confidence interval using SPSS. The result indicates that there are 95 chances out of a 100 that the population mean is somewhere between 14.68 and 14.86 years.

Notice that the range of values comprising the confidence interval is small (from only 14.68 to 14.86). This is due to the relatively large sample size of the 2021 GSS. There are 3,966 valid responses on the EDUC variable. By increasing the sample size, confidence intervals become more precise. While we will only use the standard 95% confidence interval in this course, it is possible to change the confidence level and margin of error. For example, a 99% confidence interval reduces the risk of being incorrect to only 1%. However, that reduced risk results in a less precise estimate since the interval width increases when we switch from a 95% to 99% confidence level.

Calculating confidence intervals in SPSS is easy. We use the Analyze > Descriptive Statistics > Explore command and simply move the EDUC variable into the Dependent List box and hit OK.

Exhibit 6.4. Analyze > Descriptive Statistics > Explore Output for EDUC

Explore

Descriptives

			Statistic	Std. Error
educ highest year of school completed	Mean		14.77	.044
	95% Confidence Interval for Mean	Lower Bound	14.68	
		Upper Bound	14.86	
	5% Trimmed Mean		14.82	
	Median		15.00	
	Variance		7.842	
	Std. Deviation		2.800	
	Minimum		0	
	Maximum		20	
	Range		20	
	Interquartile Range		4	
	Skewness		-.502	.039
	Kurtosis		1.686	.078

At the top of the output in Exhibit 6.4, SPSS provides the values of the mean and the standard error of the mean. The lower bound of the confidence interval is the result of the product of Z (1.96) and the standard error (.044) being subtracted from the mean (14.77). The upper bound is that product being added to the mean. The proper fashion to report this confidence interval was seen in Exhibit 6.3: $14.68 \leq \mu \leq 14.86$. We are 95% confident that the true population mean for EDUC is somewhere between 14.68 and 14.86 years.

Confidence intervals are our first engagement with inferential statistics. They allow us to make predictions about the population mean of a variable using sample data. In Chapters 2 through 4 we used the split file command in SPSS to investigate group differences in frequency distributions, measures of central tendency, and measures of dispersion. The most powerful use of confidence intervals is in comparing group means on a variable. **Statistical difference** is the idea that we are confident that group differences discovered in samples can be generalized to their larger populations. The key when working with multiple confidence intervals is to determine whether or not those intervals overlap. If they do not, we have evidence of statistical difference and can conclude that one group has a higher or lower mean than the other. When confidence intervals do overlap, this indicates that there is no statistical difference between the groups. The true population means can fall anywhere within the intervals, so we cannot be confident that any observed differences in sample means can be generalized to the larger populations when the intervals overlap.

Let's consider a fictious example. Suppose that we surveyed residents of three different cities about their weekly television viewing habits. We then calculated the three confidence intervals presented in Exhibit 6.5. Here, we see that there is no statistical difference in television viewing between residents of City A and B since the intervals overlap. That is, the upper bound of the City A interval is higher than the lower bound of the City B interval (drawing out a number line on a piece of paper and adding each confidence interval to it may be helpful for seeing overlap or lack thereof). However, residents in City C watch significantly more television than those in Cities A and B. We are confident of this statistical difference since the City C interval is higher than and does not overlap with the City A and B intervals.

Exhibit 6.5. Fictious Confidence Intervals of Television Viewing by City

City	*95% Confidence Interval for TV Hours*
A	$27.50 \leq \mu \leq 30.50$
B	$29.99 \leq \mu \leq 32.99$
C	$33.50 \leq \mu \leq 36.50$

Let's explore a real example in SPSS and compare years of formal schooling by race. As you can see in the GSS2021.sav datafile, respondents were asked to report their race. Their responses were coded into 16 different categories (see variables RACEACS1 through RACEACS16) and multiracial respondents chose multiple categories. The MINORITY variable is a dichotomous race measure that identifies people of color (coded "1") and whites (coded "0").

As we've done so far, we could use the split file command to produce the group output. However, an easier approach is to use the Factor List box in the Analyze > Descriptive Statistics > Explore dialog. So, run the command with the EDUC variable in the Dependent List and the MINORITY variable in the Factor List. As you can see in the output, the average years of schooling for white respondents is 14.85 (with a standard error of .048) and the average for people of color is 14.43 (with a standard error of .112). Exhibit 6.6 presents the confidence intervals formally so that we can determine if there is any statistical difference.

Exhibit 6.6. Confidence Intervals of Years of Education by Race

Race	*95% Confidence Interval for EDUC*
Whites	$14.76 \leq \mu \leq 14.95$
People of Color	$14.21 \leq \mu \leq 14.65$

We see that the interval for white respondents is higher than, and does not overlap with, the interval for people of color. Therefore, there is statistical difference and we are confident that whites have greater formal educational attainment than people of color. Sociologically, we know that whites tend to enjoy better life chances and opportunities than people of color.[2] Whites are more likely to attend high schools at which they are encouraged to attend college. Fortunately, the racial gap in educational

attainment has been shrinking as more and more people of color have been attending college.[3] Here at CSUF, the majority of students are people of color and about one-third are the first in their family to go to college.[4]

As you may recall, Chapter 5 mentioned that we will also be estimating π ("pie"), a population proportion, using a sample proportion, p. While we have only been discussing confidence intervals for sample means (scale variables), they can be constructed for categories of nominal and ordinal variables as well. One area that we regularly see these in is election predictions. News organizations and polling agencies regularly survey likely voters as elections approach. Variables concerning who you intend to vote for will always be nominal ones since candidate names are provided in text-based categories that cannot be ranked.

Exhibit 6.7 provides fictitious data from two polls, fielded weeks apart, asking likely voters whom they intend to support in an upcoming election. In Poll 1, we see that the intervals do not overlap. On that date, Candidate B was the "projected winner" since their interval is higher than that of Candidate A. In the second poll, as election day approached, we see that the confidence intervals for the two candidates do overlap. Candidate A's interval is that same as in the first poll, but the one for Candidate B fell by three percentage points in Poll 2. Thus, at this second date, the election would be "too close to call."

Exhibit 6.7. Fictitious Confidence Intervals from Two Polls on Candidate Voting Likelihood

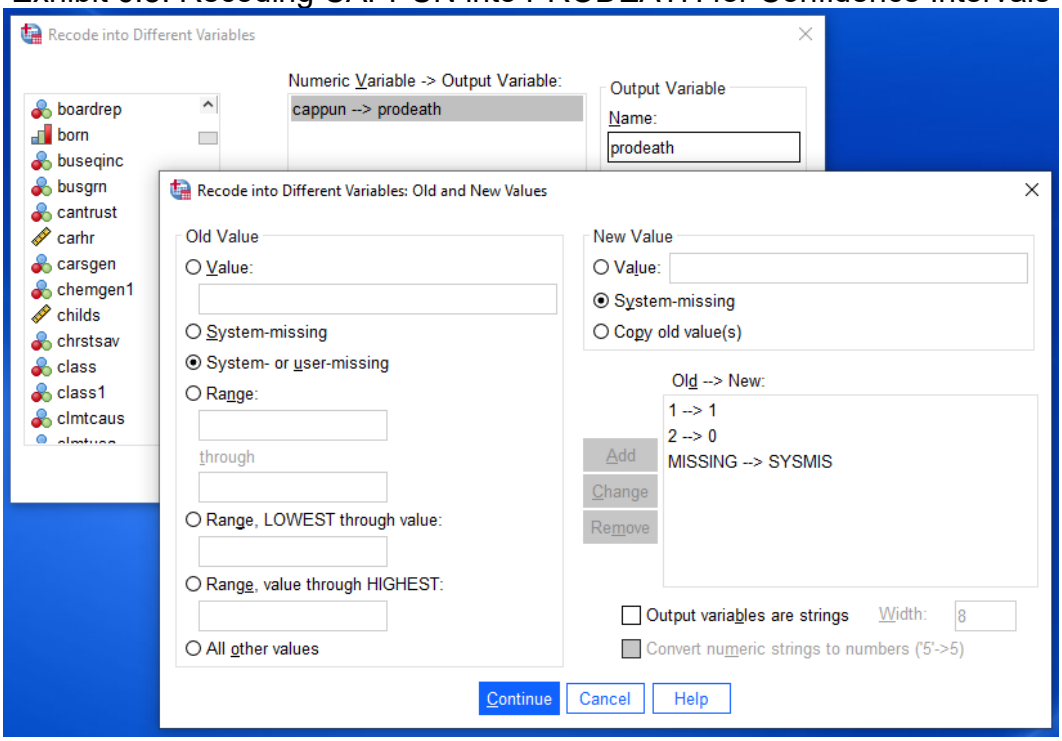
	95% Confidence Interval for Candidate A	*95% Confidence Interval for Candidate B*
Poll 1	$39.9 \leq \pi \leq 46.1$	$48.9 \leq \pi \leq 55.1$
Poll 2	$39.9 \leq \pi \leq 46.1$	$45.9 \leq \pi \leq 52.1$

So, how do we construct such confidence intervals for categories of nominal or ordinal variables? The key is that we must dummy code the variable of interest into one or more new variables with just two categories. The categories must have values of 0 and 1 so that the proportion that results from calculating the mean of the new variable can be easily interpreted as a percentage. You may recall from Chapter 1 that

dichotomous, dummy variables distinguish between the presence and absence of a characteristic.

Let's walk through the process with an analysis of the CAPPUN variable: "Do you favor or oppose the death penalty for persons convicted of murder?" In this variable, the value label "favor" is assigned the value of 1 and "oppose" is 2. We need to recode this variable into a different one that is coded 0 and 1. Refer back to Chapter 2 as needed and recode CAPPUN into a new variable entitled PRODEATH. The old value of 1 (favor) will remain as 1 in the new variable. The old value of 2 (oppose) will become 0 in the new variable (see the completed dialog presented as Exhibit 6.8).

Exhibit 6.8. Recoding CAPPUN into PRODEATH for Confidence Intervals



Now that you have created a new variable, construct a confidence interval for PRODEATH. The output is reproduced in Exhibit 6.9. First, we see that the mean is reported as .5615 and the standard error as .00789. The

mean represents the proportion of respondents in the “favor” group (coded as 1). Notice the new variable name alludes to that group. It is useful to give your new variables meaningful names that will help you remember what they are measuring. To make the mean a percentage, we simply multiply that proportion by 100 (move the decimal two places over to the right). So, 56.15% of respondents favor the death penalty for murder. Using the values for the lower bound and upper bound, we can report our confidence interval as: $54.61 \leq \pi \leq 57.70$. That is, we are 95% confident that the true population proportion of American adults who favor the death penalty for murder is between 54.61% and 57.70%. Note that once again, this range is fairly small given the large sample size on this question (N=3,957).

Exhibit 6.9. Analyze > Descriptive Statistics > Explore
Output for PRODEATH

Explore

Descriptives

			Statistic	Std. Error
prodeath	Mean		.5615	.00789
	95% Confidence Interval for Mean	Lower Bound	.5461	
		Upper Bound	.5770	
	5% Trimmed Mean		.5684	
	Median		1.0000	
	Variance		.246	
	Std. Deviation		.49626	
	Minimum		.00	
	Maximum		1.00	
	Range		1.00	
	Interquartile Range		1.00	
	Skewness		-.248	.039
	Kurtosis		-1.939	.078

Now, let’s practice our skills in analyzing multiple confidence intervals for overlap by adding MINORITY to the Factor List field in the Explore dialog.

Exhibit 6.10 presents the confidence intervals that result. Do the two intervals overlap? They do not and it is clear that a lower proportion of people of color report favoring the death penalty for murder than whites. A clear majority of white respondents are in favor, while less than half of people of color favor it. These findings are likely due to people of color's distrust of the criminal justice system and their negative perceptions of police.[5]

Exhibit 6.10. Confidence Intervals of Favoring Death Penalty by Race

Race	*95% Confidence Interval for PRODEATH*
Whites	$57.70 \leq \pi \leq 61.13$
People of Color	$39.69 \leq \pi \leq 46.59$

This chapter has provided our first experience with inferential statistics. Point estimates from sample data are limited since we have no gauge of their accuracy. Confidence intervals specify a range of values within which the population parameter is estimated to fall. We will always use a 95% confidence level in this course and accept the 5% risk of being wrong.

When we have a scale variable, estimating the population mean (μ) is easy using the Analyze > Descriptive Statistics > Explore command. Group analyses offer the most powerful application of the technique by allowing us to determine whether there are statistical differences between two or more groups that are generalizable to the larger populations. If we have a nominal or ordinal variable, we can estimate the population proportion of a category (π). To do so, we must create a dummy variable with the category of interest coded as 1 and the remaining as 0.

In the next chapter, we continue our journey into inferential statistics by returning to the steps of the research process (Chapter 1) and learning about statistical hypothesis testing. Our first inferential test statistic is the most popular one, Chi-square.

Key Terms

Point estimate, confidence intervals, confidence level, margin of error, standard error of the mean, and statistical difference

Endnotes

1. Exhibit 6.1 is licensed CC BY-SA 4.0 by CMG Lee

2. Cho, Ryan W. and Jennie E. Brand. 2019. “Life Chances and Resources.” In *The Blackwell Encyclopedia of Sociology*, edited by George Ritzer and Chris Rojek. https://doi.org/10.1002/9781405165518.wbeosl043.pub2

3. Dyer, Shauna and Giovanni Roman-Torres. 2022. “Latina/o Postsecondary Education: Trends in Racial/Ethnic Education Gaps and the Role of Citizenship in Access to Higher Education.” *Demography* 59 (6): 2053-2078.

4. See CSUF Institutional Research Dashboard tab on Demographics at https://www.fullerton.edu/data/institutionalresearch/student/enrollments/headcountsftesbycollegeandstudentlevel.php

5. Foglia, Wanda D. and Nadine M. Connell. 2019. “Distrust and Empathy: Explaining the Lack of Support for Capital Punishment Among Minorities.” *Criminal Justice Review* 44(2): 204-230.

This 2025 work is licensed under CC BY-NC-SA 4.0

Quantitative Sociology Essentials:
An Introduction to Survey Data Analysis using IBM® SPSS®

Ed Collom, Ph.D.
Professor of Sociology
California State University, Fullerton

Chapter 7. Hypothesis Testing

In this chapter we will learn about testing hypotheses on associations between two variables. While we have already been analyzing group differences using descriptive statistics and the split file command in SPSS, it is now time to engage in formal inferential statistical analyses. The first inferential test statistic that we will use is Chi-square.

Chapter Objectives

After reading this chapter, students should be able to:

- Understand the purpose of statistical hypothesis testing
- Select the chi-square test as the appropriate one for associations between categorical variables
- Apply probability level results to make a decision about the null hypothesis

In Chapter 1, we learned about the steps of the research process. The third step, formulating the research hypothesis, will now be applied so that we can analyze survey data for **bivariate associations** (step 5) and assess whether an association that we observe between two variables in our sample can be generalized to the larger population of interest (step 6).

Statistical **hypothesis testing** allows us to evaluate hypotheses about population parameters when we only have sample statistics. There are mathematical assumptions that permit us to engage in such inferential statistics. For most tests, these assumptions are easily met when we have high-quality survey data such as the General Social Survey.

Hypotheses

There are two different types of hypotheses used in the process of hypothesis testing. The **research hypothesis** is the statement reflecting the anticipated impact that an independent (X) variable will have upon the dependent variable (Y). Research hypotheses are symbolized as H_1 ("H sub one"). In Chapter 6, you will recall that we computed confidence intervals for whites and people of color (MINORITY) on the variables EDUC (years of formal schooling) and PRODEATH (attitudes toward the death penalty for murder). If we conducted a literature review, we would find theory and previous research suggesting that the educational attainment of whites will likely be higher than that of people of color. A formal research hypothesis could then be stated as: $\mu_0 > \mu_1$. This statement predicts that the population mean of group 0 (whites) is expected to be greater than the population mean of group 1 (people of color). A literature review about racial differences in support for the death penalty would suggest the following research hypothesis: $\pi_0 > \pi_1$. This statement predicts that the population proportion of group 0 (whites) is expected to be greater than the population proportion of group 1 (people of color).

The **null hypothesis** is always a statement indicating that there is no association between the independent and dependent variables. The null hypothesis predicts no group differences or no impact of X upon Y. It is symbolized as H_0 ("H sub zero"). The null directly contradicts the research hypothesis and is expressed in population parameters. Thus, the null hypothesis for the association between MINORITY and EDUC would be stated as: $\mu_0 = \mu_1$. And, the null hypothesis for the association between MINORITY and PRODEATH would be stated as: $\pi_0 = \pi_1$.

The concept of the null hypothesis is crucial in inferential statistics and hypothesis testing. Since we employ sample data and lack complete population data, we are unable to directly test the research hypothesis since it concerns group differences in the population. Statistical hypothesis testing only allows us to test the null hypothesis of no association between the variables in the larger population. The null also serves as a convenient reminder that we are always interested in whether we can generalize differences observed in our samples to the larger populations of interest. As discussed in Chapter 5, we are not directly interested in the respondents within our samples. We are interested in their characteristics because they represent many more people in the larger population.

In statistical hypothesis testing, we always hope to **reject the null** hypothesis of no association. If we have statistical evidence to reject the null, we then can have indirect support for our research hypothesis of interest. If we **fail to reject the null** hypothesis, we must conclude that there is no association between the variables in the larger population.

Type I and Type II Errors

Since our analyses are based upon sample data, there is always the possibility of error. While the likelihood of error will be quantified, we will never really know whether the null hypothesis is true or false. There are two types of errors that can be committed in statistical hypothesis testing (Exhibit 7.1 below). The first situation is when we reject the null of no association and it turns out that the variables are not actually associated in the population. Here, we have committed a **Type I error** since we should have failed to reject the null. The second situation is when we fail to reject the null, but it turns out that the null should have been rejected. Here, we have committed a **Type II error** since there actually is an association between the variables in the population.

Exhibit 7.1. Type I and II Errors

If our decision was to…	*And the actual state of affairs in the population is…*	
	H_0 is True	**H_0 is False**
Reject H_0	Type I Error	Correct Decision
Not Reject H_0	Correct Decision	Type II Error

Chi-Square Test

The first inferential test statistic that we will cover is the most widely-used one, the **chi-square test**. Note that chi ("kai") is a statistic, not positive energy ("chee") nor tasty tea ("chai"). Mispronunciation of this term is common among students. Chi-square is the appropriate inferential technique to test for a relationship between two categorical variables. That is, if our independent variable is nominal or ordinal and our dependent variable is nominal or ordinal, chi-square is appropriate.

As with all inferential tests, chi-square is a test of the null hypothesis. The null for chi-square is a statement of **statistical independence**. That is, the variables are independent of one another and not associated in the

larger population. Our research hypothesis will state that the variables are associated in the population. In other words, the Y variable is expected to be **statistically dependent** upon the X variable.

As we will learn in Chapter 8, the chi-square test is used with bivariate tables (**crosstabulations**). These are tables that simultaneously display respondent scores to two variables at one time. The **observed frequencies** of a crosstabulation are the number of cases from a sample that fall within every possible combination of the independent and dependent variables. They are the actual number of respondents that are in category 1 of the X variable and category 1 of the Y variable, category 2 of X and category 1 of Y, and so on.

While the observed frequencies are actual observations from the two variables in the sample, the **expected frequencies** are those that we would expect to find when the null hypothesis is true and there is no association between the independent and dependent variables. While we will only use SPSS to compute chi-square in this course, the calculation is based upon a formula that compares the observed frequencies to the expected frequencies.

The value of the chi-square test statistic is always a positive one. The minimum value of the statistic is zero and there is no upper limit to the value. Zero indicates that the observed and expected frequencies are identical. That is, the variables are statistically independent and the null hypothesis is true as there is no association between the variables in the population. When the observed frequencies are considerably different from the expected frequencies, we are confident that there is an association and the null hypothesis is rejected. Inferential test statistics are said to be **statistically significant** when an association between two variables that is observed in a sample is highly likely to appear in the population.

Probability and Alpha Values

The p or **probability value** is the exact likelihood that the obtained value of an inferential test statistic is a random occurrence rather than the consequence of a real association between X and Y in the population. It is a measure of how unusual or rare our test statistic is compared with our null hypothesis. For the chi-square test statistic, the p value provides the likelihood that any statistical dependence observed in the sample is due to

chance or sampling error (recall Exhibit 6.1 and the idea of error within the tails of the sampling distribution).

The smaller the p value, the less likely that the statistical dependence we observe is due to chance. With small p values, we have more evidence that the null hypothesis should be rejected since there is an actual association between X and Y in the population. When the null is rejected, our results are deemed statistically significant.

The **alpha level** (α) is the defined level of probability at which the null hypothesis is rejected. This benchmark is typically set at the .05, .01, or .001 level. In sociology, the norm is .05, so this is the alpha that will be used exclusively in this text. We will also follow the custom of using what are referred to as **two-tailed tests**. Here, the error is nondirectional and split between each of the tails of the distribution (see Exhibit 6.1 again).

With an alpha of .05, there are no more than 5 chances in 100 that the statistical dependence we observe in a sample is due to chance. In other words, we are at least 95% confident that the statistical dependence observed in the sample reflects an actual association in the population. Since SPSS will report the p value with three decimal places, our alpha will actually be defined as .054 (which rounds down to .05). This 5% margin of error is the standard in survey research as well as sociology.

This intentionally brief chapter has attempted to present the essentials of hypothesis testing. Every inferential test statistic that we produce in SPSS will have a probability value associated with it. With our alpha benchmark of .054, we will reject the null hypothesis of no association when the p value is .054 or less ($p \leq \alpha$). In such cases, we are able to document a statistically significant association between two variables that we are highly confident exists in the population. Now, we are able to move on to Chapter 8 and engage in bivariate analyses by testing for associations between categorical variables using crosstabulations and chi-square.

Key Terms

Bivariate associations, hypothesis testing, research hypothesis, null hypothesis, reject the null, fail to reject the null, Type I error, Type II error, chi-square test, statistical independence, statistical dependence, crosstabulations, observed frequencies, expected frequencies, statistically significant, probability value, alpha value, and two-tailed tests.

This 2025 work is licensed under CC BY-NC-SA 4.0

Quantitative Sociology Essentials: An Introduction to Survey Data Analysis using IBM® SPSS®

Ed Collom, Ph.D.
Professor of Sociology
California State University, Fullerton

Chapter 8. Crosstabulations and Measures of Association

With knowledge of all of the concepts in the first seven chapters of this text, we are now prepared to compute inferential test statistics and engage in bivariate analyses. Crosstabulations are the first technique that we will explore to test for associations between categorical (nominal or ordinal) variables. These are also commonly referred to as bivariate or contingency tables. Crosstabulations are a widely used and powerful analytical tool in survey research when combined with the chi-square test statistic and measures of association.

Chapter Objectives

After reading this chapter, students should be able to:

- Understand the anatomy of a crosstabulation
- Produce a crosstabulation with the chi-square test and the appropriate measure of association to test for an association between two variables
- Evaluate hypotheses by reporting column percentages, generalizability, and strength of association

As with measures of central tendency and measures of dispersion, the levels of measurement of the two variables that we are interested in analyzing determines the appropriate statistical technique to test for associations. As seen in Exhibit 8.1, we use crosstabs in many scenarios. When our independent and dependent variables are nominal or ordinal, crosstabs are appropriate. And, as we shall see, if we happen to have an independent variable that is a scale one, we can recode it into fewer categories to make it ordinal and use crosstabs as well.

Exhibit 8.1. Level of Measurement of Variables
Determines Bivariate Statistical Test

DEPENDENT VARIABLE	INDEPENDENT VARIABLE: *Nominal or ordinal with two categories*	*Nominal or ordinal with 3 or more categories*	*Scale*
Nominal or ordinal with two categories	Crosstab	Crosstab	Recode independent variable to use crosstab
Nominal or ordinal with 3 or more categories	Crosstab	Crosstab	Recode independent variable to use crosstab
Scale	Independent samples t test	Analysis of variance (ANOVA)	Regression techniques

Crosstabulations

A crosstabulation is a table that simultaneously displays respondent scores to two variables at one time. When constructing a crosstabulation in SPSS, we are required to identify the column variable and the row variable. As seen in Exhibit 8.2, the independent variable (X) goes in the columns of the crosstabulation while the dependent variable (Y) belongs in the rows. It is crucial to generate crosstabulations correctly in SPSS. Otherwise, the results will not be accurate.

The observed frequencies of a crosstabulation will be found within the intersections of the columns and rows. These **cells** (see area in yellow) contain the number of cases from a sample that fall within every possible combination of the independent and dependent variables. The example in Exhibit 8.2 is the skeleton of what is referred to as a 2x2 crosstab in that both the independent and dependent variables have two categories. If the X variable had 3 categories and Y had 2, it would be a 3x2 table.

Crosstabulations also have a row total (last column of the table) and a column total (last row of the table). These are also referred to as marginals. The grand total, or number of cases analyzed in the table, is always found in the lower right-hand corner. The row total is a frequency distribution of the dependent variable and the column total is a frequency distribution of the independent variable.

Exhibit 8.2. Anatomy of a Crosstabulation

DEPENDENT VARIABLE	INDEPENDENT VARIABLE: *Category 1*	*Category 2*	*Row Total*
Category 1			
Category 2			
Column Total			Total Cases (N)

The sample in Exhibit 8.3 adds more detail. Some sociologists use survey data to study the predictors of social movement or protest participation.[1] Suppose we had a sample of 50 males and 50 females who reported whether or not they had ever attended a protest event. To test for sex differences in protest participation, we would place the independent variable (SEX) in the columns and the dependent variable (PROTEST) in the rows. While the cells (yellow area) are still blank, we see in this simple example that 50 of the respondents reported that they had attended a protest event in the past (yes) and 50 reported that they had not (no).

Exhibit 8.3. Sample Crosstabulation of Protest Participation by Sex with Empty Cells

		SEX		
		Male	*Female*	*Row Total*
PROTEST PARTICIPATION	*Yes*			50
	No			50
	Column Total	50	50	100

Exhibit 8.4 adds in some fictitious frequencies for this example. Here, we see that all 50 female respondents reported that they had previously participated in a protest event whereas all 50 of the male respondents reported that they had not. This is an example of a **perfect association** between the two variables. All of the females are in the "yes" group and all of the males are in the "no" group. In this case, one's sex perfectly determines their protest participation. With knowledge of the score on X, we would know the score on Y.

Exhibit 8.4. Fictitious Crosstabulation of Protest Participation by Sex (Perfect Association)

PROTEST PARTICIPATION	SEX: *Male*	SEX: *Female*	*Row Total*
Yes	0	50	50
No	50	0	50
Column Total	50	50	100

Exhibit 8.5. Fictitious Crosstabulation of Protest Participation by Sex (Perfect Nonassociation)

PROTEST PARTICIPATION	SEX: *Male*	SEX: *Female*	*Row Total*
Yes	25	25	50
No	25	25	50
Column Total	50	50	100

What do you think a crosstabulation of a **perfect nonassociation** would look like? If you were thinking that the cases would be evenly dispersed across the four cells of this 2x2 crosstab, you would be correct. As the example in Exhibit 8.5 illustrates, the 100 cases would be evenly distributed with 25 cases in each cell. This indicates that one's sex has no influence upon one's protest participation. In other words, knowledge of X does not improve our ability to predict Y.

In these simple examples, it is easy to ascertain relationships in 2x2 tables with 100 cases (a nice, round number). As you can imagine, actual crosstabs of variables from survey data can have many categories containing thousands of cases. While we have been using frequencies in the cells so far, we will actually need to use percentages in these tables to ascertain whether associations exist. In SPSS, we will always request column percentages. Once our crosstab is produced by the software, we must confirm that these are present. **Column percentages** always sum up to 100% in the Column Total row. If the percentages in the last row of your crosstab are not values of 100% across, it has not been properly generated.

Exhibit 8.6 adds in the column percentages to our last example. Now, the cells contain both the frequency and the column percent. We report the findings from our crosstabs by comparing the percentages within the rows, across the columns. That is, we compare the percentages across the different categories of the independent variable. Percentage differences within the rows of a crosstab indicate that the X variable is producing an impact upon the Y variable. In this example, there is no association since 50% of male respondents and 50% of female respondents report previous protest participation. And, 50% of male respondents and 50% of female respondents report no previous protest participation. If we added column percentages to the crosstab in Exhibit 8.4, we would see that 0% of males and 100% of females report having previously participated in a protest event. In reporting the results of a crosstab with column percentages, it is critical to place the X variable in the columns, the Y variable in the rows, and to compare the percentages from the categories of X to a category of Y one row at a time.

Exhibit 8.6. Fictitious Crosstabulation of Protest Participation by Sex with Column Percentages (Perfect Nonassociation)

		SEX		
		Male	*Female*	*Row Total (%)*
PROTEST PARTICIPATION	*Yes (f)* *(%)*	25 50%	25 50%	50 50%
	No (f) *(%)*	25 50%	25 50%	50 50%
	Column Total *(%)*	50 100%	50 100%	100 100%

Are we ready to get into SPSS and produce our very first crosstabulation? I think we are! Open the GSS2021.sav data file. As you may recall from Chapter 6, we used confidence intervals to explore racial differences in the CAPPUN variable (favoring or opposing the death penalty for those convicted of murder). Now, let's use the more powerful technique of crosstabulations. Since MINORITY is a nominal variable and CAPPUN is an ordinal one, crosstabulations are the appropriate technique to test for a bivariate association here (refer back to Exhibit 8.1). The command is Analyze > Descriptive Statistics > Crosstabs. Place the dependent variable CAPPUN in the Row(s) box and the independent variable MINORITY in the Column(s) box. Then, click the Cells button on the right side. We will always have to use the Cells dialog and the only thing that we will ever want to do in there is to check off Column under the Percentages box on the left side (see Exhibit 8.7). Then, click Continue and OK to generate the crosstab.

Exhibit 8.7. Cells Dialog for Crosstab of CAPPUN by MINORITY

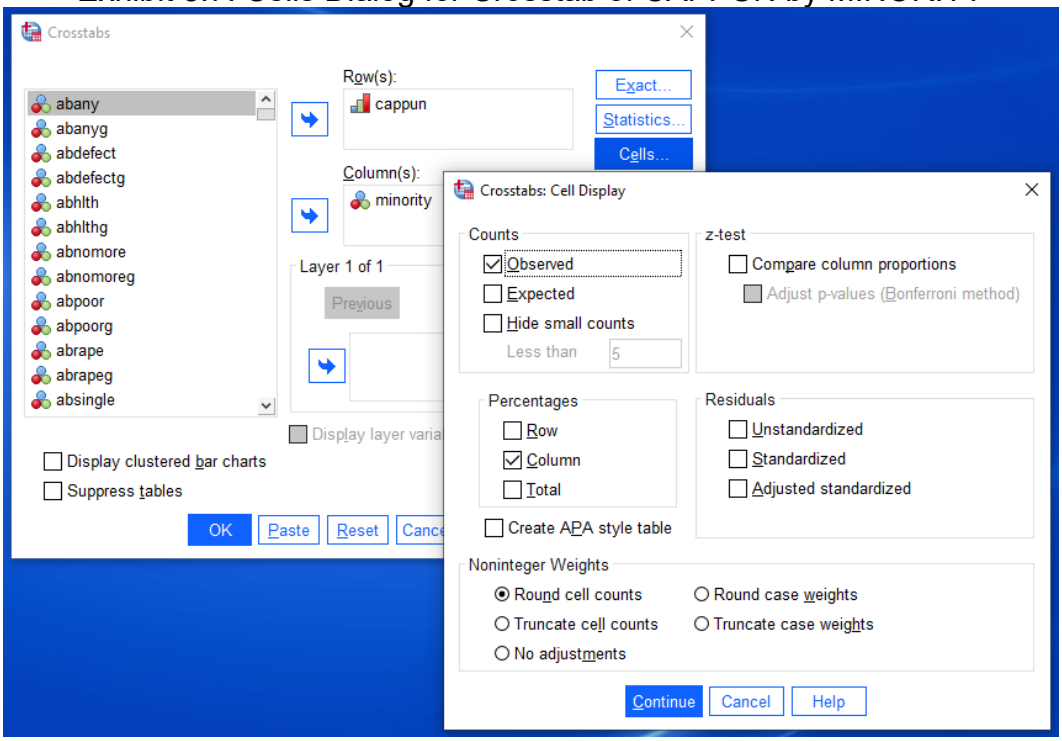


Exhibit 8.8 displays the SPSS output. As we can see in the lower right corner, this crosstab is based on 3,944 valid cases. The Row Total (last column) indicates that overall, 2,214 respondents (56.1%) “favor” the death penalty for murder while 1,730 respondents (43.9%) “oppose” it. The Count of the Column Total (last row) shows that 3,149 respondents identify as white and 795 respondents identify as a person of color. Since we requested column percentages, those totals will always be 100% (59.4% + 40.6% = 100% and 43.1% + 56.9% = 100%).

Exhibit 8.8. SPSS Output for Crosstabulation of CAPPUN by MINORITY

Crosstabs

cappun favor or oppose death penalty for murder * minority people of color (dichotomous race) Crosstabulation

			minority people of color (dichotomous race)		
			0 White	1 Person of Color	Total
cappun favor or oppose death penalty for murder	1 favor	Count	1871	343	2214
		% within minority people of color (dichotomous race)	59.4%	43.1%	56.1%
	2 oppose	Count	1278	452	1730
		% within minority people of color (dichotomous race)	40.6%	56.9%	43.9%
Total		Count	3149	795	3944
		% within minority people of color (dichotomous race)	100.0%	100.0%	100.0%

The key to reporting crosstabulation findings is to focus on the column percentages within the cells and to make comparisons within the rows, across the columns. 59.4% of white respondents favor the death penalty for murder compared to only 43.1% of people of color. 40.6% of whites oppose the death penalty while 56.9% of people of color are in opposition. Remember that differences in these percentages reflect an association between the variables. In other words, X (MINORITY) appears to have some influence upon Y (CAPPUN). You may be wondering what magnitude of percentage difference is required to suggest an association. In a 2x2 crosstab such as this one, the rule of thumb is that we would need to see at least a five percentage point difference for an association to be implied. As we shall see later in this chapter, there are statistics that we will add to clarify the existence and strength of associations in crosstabs.

Let's run another one. Back in Chapter 4, we used the split file command to explore sex differences in attitudes about pornography laws. Both SEXBIRTH1 and PORNLAW are nominal variables. Produce the crosstab in SPSS and confirm that your output matches that in Exhibit 8.9.

Exhibit 8.9. SPSS Output for Crosstabulation
of PORNLAW by SEXBIRTH1

Crosstabs

pornlaw feelings about pornography laws * sexbirth1 r's sex assigned at birth (2021) Crosstabulation

			sexbirth1 r's sex assigned at birth (2021)		
			1 male	2 female	Total
pornlaw feelings about pornography laws	1 there should be laws against the distribution of pornography whatever the age	Count	191	475	666
		% within sexbirth1 r's sex assigned at birth (2021)	16.9%	32.6%	25.8%
	2 there should be laws against the distribution of pornography to persons under 18	Count	858	930	1788
		% within sexbirth1 r's sex assigned at birth (2021)	76.1%	63.9%	69.2%
	3 there should be no laws forbidding the distribution of pornography	Count	78	51	129
		% within sexbirth1 r's sex assigned at birth (2021)	6.9%	3.5%	5.0%
Total		Count	1127	1456	2583
		% within sexbirth1 r's sex assigned at birth (2021)	100.0%	100.0%	100.0%

16.9% of male respondents selected the “there should be laws against the distribution of pornography whatever the age” response while 32.6% of females support restricting distribution of pornography. There is almost a two-fold difference in these percentages, indicating that there certainly appears to be an association between these variables. Females are more likely to support restriction and are less likely to believe that there should be no restrictions (see the last row for category 3; 6.9% vs. 3.5%).

When crosstabulations are comprised of two ordinal variables, we can determine if an association is positive or negative in direction. Since ordinal variables are ranked, we can see if higher values on the independent variable correspond with higher values on the dependent variable (a **positive association**). When higher values on X correspond with lower values on Y, there is a **negative association**. We’ve already looked at both the DEGREE and POLVIEWS variables, so let’s see if education attainment predicts political ideology. Produce the crosstab and confirm that it is the same as Exhibit 8.10.

Exhibit 8.10. SPSS Output for Crosstabulation of POLVIEWS by DEGREE

Crosstabs

polviews think of self as liberal or conservative * degree r's highest degree Crosstabulation

			degree r's highest degree					
			0 less than high school	1 high school	2 associate/juni or college	3 bachelor's	4 graduate	Total
polviews think of self as liberal or conservative	1 extremely liberal	Count	10	67	8	60	61	206
		% within degree r's highest degree	4.5%	4.3%	2.2%	5.8%	8.1%	5.2%
	2 liberal	Count	20	159	44	206	192	621
		% within degree r's highest degree	9.0%	10.1%	12.0%	20.1%	25.4%	15.7%
	3 slightly liberal	Count	20	141	38	159	131	489
		% within degree r's highest degree	9.0%	9.0%	10.4%	15.5%	17.3%	12.4%
	4 moderate, middle of the road	Count	104	635	153	297	178	1367
		% within degree r's highest degree	46.8%	40.4%	41.8%	28.9%	23.5%	34.7%
	5 slightly conservative	Count	21	199	45	119	90	474
		% within degree r's highest degree	9.5%	12.7%	12.3%	11.6%	11.9%	12.0%
	6 conservative	Count	33	279	57	158	87	614
		% within degree r's highest degree	14.9%	17.7%	15.6%	15.4%	11.5%	15.6%
	7 extremely conservative	Count	14	92	21	27	18	172
		% within degree r's highest degree	6.3%	5.9%	5.7%	2.6%	2.4%	4.4%
Total		Count	222	1572	366	1026	757	3943
		% within degree r's highest degree	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Notice that this 5x7 crosstab is a large one with 35 cells. When reporting the findings, sociologists usually focus on the rows with the largest percentage differences to highlight any association. In this one, the "liberal" row best shows the trend. 9.0% of those who did not graduate high school identify as liberal, 10.1% of high school graduates identify as liberal, 12.0% of those with junior college degrees identify as liberal, 20.1% of those with Bachelor's degrees identify as liberal, and 25.4% of those who have earned graduate degrees identify as liberal. Notice that the percentages increase across the degree groups. Those with greater educational attainment are more likely to identify as liberal.[3]

The percentages within the "extremely conservative" row decrease as educational attainment increases: 6.3% of those who did not graduate high school identify as extremely conservative whereas only 2.4% of those with graduate degrees identify as such. When we think about direction to an association between ordinal variables in a crosstab, we must consider how the variables are coded. The X variable is coded from low to high educational attainment. The Y variable is coded from "extremely liberal" (1) to "extremely conservative" (7). When the percentage values within the last row decrease as they do here, we have evidence of a negative

association. In other words, those with greater educational attainment are less likely to identify as "extremely conservative." Overall, respondents with higher scores on X tend to have lower scores on Y.

Chi-Square and Measures of Association

Now, we need to integrate the **chi-square** test statistic into our crosstabulation analyses. As you will recall from the previous chapter, chi-square is the appropriate inferential technique to test for a relationship between two categorical variables. As with all inferential tests, chi-square is a test of the null hypothesis. The null for chi-square is a statement that there is no association or statistical independence between X and Y. When there is no association in a crosstabulation, the column percentages will be of very similar values within the rows.

Every inferential test statistic that we produce in SPSS will have a probability value associated with it. Inferential test statistics are said to be statistically significant when an association that is observed between two variables in a sample is highly likely to appear in the population. With our alpha benchmark of .054, we will reject the null hypothesis of no association when the p value of chi-square is .054 or less ($p \leq \alpha$). In such cases, we are able to document a statistically significant association between two variables in a crosstabulation that we are highly confident exists in the population.

The chi-square test is a critical component of a crosstabulation. Sociologists and survey researchers typically begin by looking at chi-square to determine if there is an observed association that can be generalized. If there is not, we often do not even bother looking for column percentage differences since we cannot reject the null.

As alluded to earlier, sociologists are also interested in quantifying the strength of an observed association between an independent and dependent variable. **Measures of association** are summary statistics that quantify the strength, and if applicable, the direction of a relationship between an independent and dependent variable (psychologists refer to these as effect size).

As with all of the statistics that are employed in this text, the level of measurement of the variables determines which measure of association to use. While several of these exist for various variable types, we will make

life easier here and only focus on the most commonly used measures of association. Thus, if a crosstabulation contains a nominal variable, the appropriate measure of association is **Cramer's V**.

Cramer's V is a **nondirectional** measure of association and has a convenient metric ranging from 0.00 to 1.00. With nominal variables, the response categories cannot be ranked. Therefore, it does not make sense to think about an association involving a nominal variable as being directional. A value of zero signifies no association (perfect nonassociation) while a value of one indicates a perfect association. However, it is rare that our Cramer's V will be exactly 0.00 or 1.00, it almost always falls somewhere in between. Quantitative sociologists prefer to use descriptive terms such as "weak," "moderate," or "strong" when referring to the strength of associations. Thus, Exhibit 8.11 presents the **strength guidelines** that will be employed in this text.[2]

Exhibit 8.11. Strength Guidelines for Measures of Association

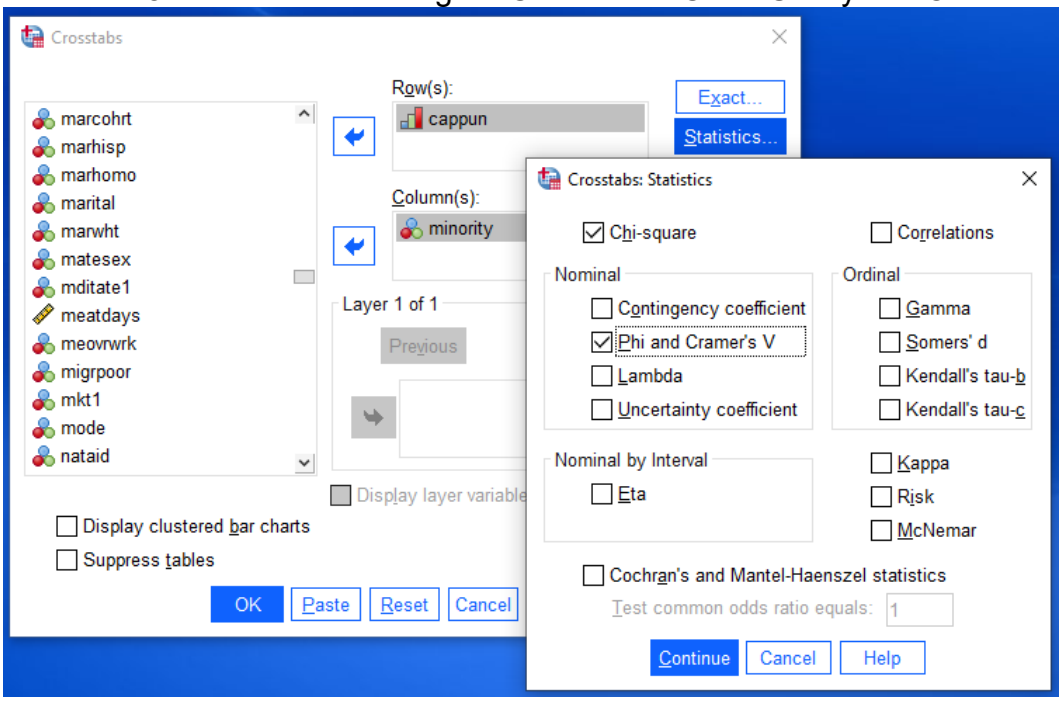
Value of Statistic	Strength of Association
0.00	None (no association)
± .015 to .194	Weak
± .195 to .394	Moderate
± .395 to .999	Strong
± 1.00	Perfect

We will see where these values appear in SPSS shortly. When our crosstabulations contain ordinal variables (or the rare instance of an ordinal by a scale one), we will use **Gamma** as the appropriate measure of association. Gamma is a **directional** measure (since ordinal variables are ranked). It ranges from -1.00 to +1.00 with -1.00 indicating a perfect negative association and +1.00 reflecting a perfect positive relationship. The guidelines in Exhibit 8.11 are stated with the ± sign since Gamma can be positive or negative. Notice that the direction does not impact the strength, these are separate things.

Now that we know that our crosstabulations need to be complemented with chi-square and the appropriate measure of association, let's revisit the three crosstabs that we produced earlier. Exhibit 8.8 displays CAPPUN by MINORITY. We can now reproduce this and use the

Statistics dialog to get chi-square and the appropriate measure of association (see Exhibit 8.12).

Exhibit 8.12. Statistics Dialog for Crosstab of CAPPUN by MINORITY



In the Statistics dialog we need to request chi-square (top left) and Cramer's V as our appropriate measure of association since the MINORITY variable is a nominal one. Under the Nominal box on the left side, check off Phi and Cramer's V (we will ignore the value of Phi). Now, click Continue and OK.

Exhibit 8.13 provides the statistics that are produced. SPSS always generates more output than what we will need. The statistics that are circled are the only ones we need to focus upon. First, we need to determine if chi-square is statistically significant. Remember that we will reject the null hypothesis of no association when the p value of chi-square is .054 or less ($p \leq \alpha$). The p value is found under the Asymptotic Significance column and is reported as .001. This is clearly below our alpha of .054 and thus, we can reject the null of no association. In other words, we are highly confident that there is an association between MINORITY and CAPPUN that is generalizable to the larger population.

What is the strength of this statistically significant association? The value of Cramer's V is reported as .132 which indicates a "weak" association per our strength guidelines. Thus, there is a weak and statistically significant association between the variables. Do not get overly concerned that an association is in the weak range. If an association is statistically significant, it is sociologically interesting. We are engaging in bivariate analyses at this point and there are certainly factors other than race that will influence attitudes toward the death penalty too. Also, remember that we are studying people's characteristics, attitudes, and behaviors. Sociologists understand that people are usually not very consistent nor predictable.

Exhibit 8.13. SPSS Output for Chi-Square and Measures of Association

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	68.247[a]	1	<.001		
Continuity Correction[b]	67.588	1	<.001		
Likelihood Ratio	67.786	1	<.001		
Fisher's Exact Test				<.001	<.001
Linear-by-Linear Association	68.230	1	<.001		
N of Valid Cases	3944				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 348.72.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.132	<.001
	Cramer's V	.132	<.001
N of Valid Cases		3944	

Now, let's revisit our PORNLAW by SEXBIRTH1 analysis in Exhibit 8.9 by producing chi-square and the appropriate measure of association. For the latter, we will use Cramer's V once again since both of these variables are nominal ones. What did you find? Confirm for yourself that there is a

weak (Cramer's V = .186) and statistically significant (p value of chi-square = .001) association between the variables.

In Exhibit 8.10, we analyzed POLVIEWS by DEGREE. Both of these variables are measured at the ordinal level, so Gamma is our appropriate measure of association (found under the Ordinal box on the right side of the Statistics dialog). Reproduce the analysis and confirm that there is a negative, moderate association (Gamma = -.197) that is highly statistically significant (p value of chi-square = .001). Respondents with greater educational attainment tend to be more liberal (they have lower scores on the dependent variable).

Back in Chapter 3, we explored differences in religious service attendance by marital status groups. Let's continue to build our skills and crosstab ATTEND by MARITAL. As you will see, this is a large table (5x9) with 45 cells. The "every week" row contains the largest percentage differences: 15.7% of married respondents and 23.3% of widowed respondents report attending religious services every week. On the other hand, only 9.9% of divorced respondents, 10.9% of separated respondents, and 8.3% of respondents who have never been married report weekly attendance. The association is weak (Cramer's V = .107) and highly statistically significant (p value of chi-square = .001).

What about ATTEND by DEGREE? This is a quick one since we cannot reject the null hypothesis of no association. Chi-square is not statistically significant (p value = .376). Also, notice how similar the column percentages are within each row. Educational attainment does not influence religious service attendance.

Let's see if DEGREE has an impact upon attitudes toward same sex relationships. Since 1973, the General Social Survey has fielded the HOMOSEX question: "What about sexual relations between two adults of the same sex – do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all?" This is an odd question that is not well-worded. It is a leading question which implies that same sex relations are wrong as the response categories measure wrongness. However, as mentioned earlier, the GSS sociologists are often reluctant to revise questions since comparability to earlier years would be lost. Exhibit 8.14 provides the SPSS output including the chi-square test results as well as Gamma (the appropriate measure of association since both variables are ordinal).

Exhibit 8.14. SPSS Output for Crosstabulation of HOMOSEX by DEGREE

homosex homosexual sex relations * degree r's highest degree Crosstabulation

			degree r's highest degree					
			0 less than high school	1 high school	2 associate/junior college	3 bachelor's	4 graduate	Total
homosex homosexual sex relations	1 always wrong	Count	68	313	80	148	82	691
		% within degree r's highest degree	44.4%	31.1%	32.4%	21.5%	16.4%	26.6%
	2 almost always wrong	Count	10	49	13	23	15	110
		% within degree r's highest degree	6.5%	4.9%	5.3%	3.3%	3.0%	4.2%
	3 wrong only sometimes	Count	8	73	13	45	41	180
		% within degree r's highest degree	5.2%	7.2%	5.3%	6.5%	8.2%	6.9%
	4 not wrong at all	Count	67	572	141	473	363	1616
		% within degree r's highest degree	43.8%	56.8%	57.1%	68.7%	72.5%	62.2%
Total		Count	153	1007	247	689	501	2597
		% within degree r's highest degree	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	92.364[a]	12	<.001
Likelihood Ratio	92.938	12	<.001
Linear-by-Linear Association	79.579	1	<.001
N of Valid Cases	2597		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 6.48.

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Gamma	.237	.026	8.860	.000
N of Valid Cases		2597			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

As we can see, the chi-square is statistically significant. The null hypothesis is rejected as there is an association between the variables that is generalizable. The row total column provides the frequency distribution of the dependent variable. More than six out of ten respondents (62.2%) chose the “not wrong at all” response in 2021. Around one-quarter (26.6%) chose the “always wrong” response. Notice that this question is polarizing as not many chose “almost always wrong” (4.2%) nor “wrong only sometimes” (6.9%).

The percentages in the first row show a trend of decline as 44.4% of those who did not graduate high school feel that same sex relations are always wrong while only 16.4% of graduate school degree holders felt the same. The last row demonstrates a steady increase of the percentages across

the degree groups: 43.8% of those who did not graduate from high school feel same sex relations are “not wrong at all” while 72.5% of graduate school degree holders feel similarly. The column percentages in the last row of an ordinal by ordinal crosstab will indicate the direction of the association. When they increase, the association is positive. When the percentages decrease, the association is negative. The value of Gamma confirms that the association is positive and moderate in strength. Those with greater educational attainment are more likely to believe that same sex relations are “not wrong at all.”

Exhibit 8.15. SPSS Output for Crosstabulation of CONMEDIC by CLASS

conmedic confidence in medicine * class subjective class identification Crosstabulation

			class subjective class identification				
			1 lower class	2 working class	3 middle class	4 upper class	Total
conmedic confidence in medicine	1 a great deal	Count	75	306	617	70	1068
		% within class subjective class identification	29.3%	30.0%	48.7%	62.5%	40.2%
	2 only some	Count	140	597	568	38	1343
		% within class subjective class identification	54.7%	58.5%	44.8%	33.9%	50.6%
	3 hardly any	Count	41	117	82	4	244
		% within class subjective class identification	16.0%	11.5%	6.5%	3.6%	9.2%
Total		Count	256	1020	1267	112	2655
		% within class subjective class identification	100.0%	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	131.335[a]	6	<.001
Likelihood Ratio	131.331	6	<.001
Linear-by-Linear Association	112.187	1	<.001
N of Valid Cases	2655		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.29.

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Gamma	-.327	.028	-11.309	<.001
N of Valid Cases		2655			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The last crosstabulation that we will consider is CONMEDIC by CLASS (Exhibit 8.15). GSS respondents were asked how much confidence they have in the people running the institution of medicine (a timely question in 2021 given the pandemic). They were also asked to identify which social class they feel they belong to. Both of the variables are ordinal, so once again, we will use Gamma as our measure of association.

The chi-square is statistically significant. 29.3% of those who identify as lower class, 30.0% of the working class, 48.7% of the middle class, and 62.5% of the upper class have a great deal of confidence in medicine. The last row ("hardly any") shows declining percentages which indicate a negative association. The value of Gamma is indeed negative and moderate in strength. As social class increases, people are less likely to have low confidence in the institution of medicine.

Several of the examples in this chapter illustrate the importance of paying close attention to how ordinal variables are coded. It is always easiest if these variables range from low to high. Sociologists often recode ordinal variables and flip the scale so that crosstabulation percentages and the direction of Gamma make more sense. In the previous example, it would be more sensible to report that as social class increases, people are more likely to have greater confidence in medicine (rather than less likely to have low confidence).

As evident, crosstabulations are a powerful analytical test of the association between two variables in a sample. However, the technique does have limitations. Large tables with many cells are often difficult to interpret. More importantly, the chi-square test statistic becomes unreliable when there are too few cases in some of the cells (the rule of thumb is a minimum of 5 expected frequencies per cell). Scale variables typically have far too many categories to use in a crosstabulation. We can always recode a variable such as AGE into an ordinal one with a handful of age categories for use in a crosstab. However, this does create other dilemmas. How may categories should there be? What are the cutpoints for such categories? Quantitative sociologists must justify such decisions and typically engage theoretical literatures or previous research to do so. In the next chapter we learn a technique to test for group differences on dependent variables measured at the scale level.

Key Terms

Cells, perfect association, perfect nonassociation, column percentages, positive association, negative association, chi-square, measures of association, Cramer's V, nondirectional, strength guidelines, Gamma, and directional.

Endnotes

1. For examples, see a): Oser, Jennifer. 2022. "Protest as One Political Act in Individuals' Participation Repertoires: Latent Class Analysis and Political Participant Types." *American Behavioral Scientist* 66 (4): 510-532. b) Caren, Neal, Raj Andrew Ghosal, and Vanesa Ribas. "A Social Movement Generation: Cohort and Period Trends in Protest Attendance and Petition Signing." American Sociological Review 76 (1): 125-151.

2. For other strength of association guidelines, see a) Frankport-Nachmias, Chava, Anna Leon-Guerrero, and Georgiann Davis. 2021. *Social Statistics for a Diverse Society* 9ed. Thousand Oaks, CA: Sage Publications. b) Kendrick, J. Richard. 2004. *Social Statistics: An Introduction Using SPSS for Windows* 2ed. Upper Saddle River, NJ: Pearson Education.

3. See Weakliem, David L. 2002. "The Effects of Education on Political Opinions: An International Study." *International Journal of Public Opinion Research* 14 (2): 141-157.

This 2025 work is licensed under CC BY-NC-SA 4.0

Quantitative Sociology Essentials:
An Introduction to Survey Data Analysis using IBM® SPSS®

Ed Collom, Ph.D.
Professor of Sociology
California State University, Fullerton

Chapter 9: *t* Testing

As seen in Exhibit 8.1 in the previous chapter, the level of measurement of the variables that we are analyzing determines the appropriate bivariate statistical test to employ. The remaining three chapters of this text cover techniques to analyze associations involving dependent (Y) variables with a level of measurement of scale. This chapter focuses exclusively on *t* testing for differences in sample means.

Chapter Objectives

After reading this chapter, students should be able to:

- Determine with which variable types the *t* test is the appropriate inferential test statistic
- Produce an independent samples *t* test to test for an association between two variables
- Evaluate hypotheses by reporting generalizability and group mean scores

t is the inferential test statistic that is used to test the null hypothesis of no difference between two means in a sample. To use *t*, the dependent (Y) variable must be scale. The independent (X) variable can be nominal or ordinal, its level of measurement does not matter. The SPSS dialog is referred to as the **independent samples *t* test**. The idea of independent samples takes us back to Chapter 1 and the concept of mutual exclusiveness, the fact that the response categories of a variable cannot overlap. Independent samples can be thought of as two groups that do not overlap.

Fundamentally, *t* testing allows us to determine if any differences in means observed in two groups of respondents (samples) can be generalized to the larger populations. The major limitation of the independent samples *t* test is that it can only test one comparison between two groups. The classic example is testing for sex differences to see if males and females differ on some trait, behavior, or attitude. While our independent (X) variable need not be a dichotomous one, we must choose the two categories to compare as we can only test for differences between two sample means.

t Testing in SPSS

The command in SPSS is Analyze > Compare Means and Proportions > Independent-Samples T Test. In the dialog, we must identify the **Test Variable** which is the dependent (Y) variable measured at the scale level. The **Grouping Variable** is the independent (X) variable. When a variable is moved into that box, you will notice that "(? ?)" appears next to the variable name. When a question mark appears in SPSS, it is alerting the user that it requires more information. In this case, we must click the **Define Groups** button to open that dialog and specify which two categories of the independent variable we want to compare. Remember that all response categories to variables in SPSS have been assigned values. So, we will have to enter the values of the two categories that we wish to compare. The default values of "1" and "2" will automatically be filled in. Those will often have to be changed, however. After the groups are defined, we will click Continue. At the bottom of the Independent-Samples T Test dialog, you will notice that there will be a check next to "Estimate effect sizes." We will not focus on that output in this text, so that checkmark can be removed or left alone and the resulting output ignored.

SPSS automatically calculates multiple types of *t* tests, so there is some output to which we will not pay any attention. For example, SPSS generates both one-tailed and two-tailed tests. We will always follow the norm and use two-tailed tests in this text, so we will be able to ignore the one-sided output. Remember that our alpha is set at .054. Therefore, we will be looking for p values or "significance levels" of .054 or less.

The default independent samples *t* test assumes that there is **equality of variances**. That is, the test is calculated with the assumption that the dispersion of the responses to the test variable is similar within both groups. In other words, the *t* formula that SPSS is computing assumes

that the standard deviations of the two groups are nearly identical. If the dispersion is not similar, then an adjusted *t* test formula that is more appropriate is used by the software.

Levene's test for the equality of variances is included in the output so that we know which *t* test results to focus upon. Levene's test is a statistic testing the null hypothesis that the variance in Y across the two categories of X is equal. The test is reported as F in SPSS. As with other tests, the value of F is not of direct importance to us, the key is whether it is statistically significant or not. The smaller the p value of the test, the less likely it is that a difference in the variances found in the samples is due to chance. Thus, if F is statistically significant (Sig. ≤ .054), then the variances are not equal. In these cases, we will refer to the bottom row of the SPSS output for the *t* test labeled, "Equal variances not assumed." If F is not statistically significant (Sig. > .054), then the variances are to be considered equal and we will be referring to the top row, "Equal variance assumed."

Once we know which row of the *t* test results is appropriate for our analysis, we are able to quickly determine if the difference in the sample means is generalizable to the larger populations. The only other portion of the remaining output that we need to pay attention to is the "Two-Sided p" column. If we see that the appropriate p value is less than or equal to .054, we are able to reject the null of no difference and conclude that the difference observed in the sample means does apply to the population as well. In instances in which the p value is greater than .054, we fail to reject the null and conclude that there are no statistical differences between the means in the population.

Let's look at the first of five examples that we will explore in this chapter. Is one's race associated with the frequency that one watches television? Open the GSS2021.sav data file and navigate to Analyze > Compare Means and Proportions > Independent-Samples T Test. The Test Variable is TVHOURS, a scale one ranging from 0 to 24 hours. The Grouping Variable is MINORITY, a dichotomous nominal variable with whites coded "0" and people of color coded "1."

Exhibit 9.1. SPSS Define Groups Dialog of Independent-Samples T Test

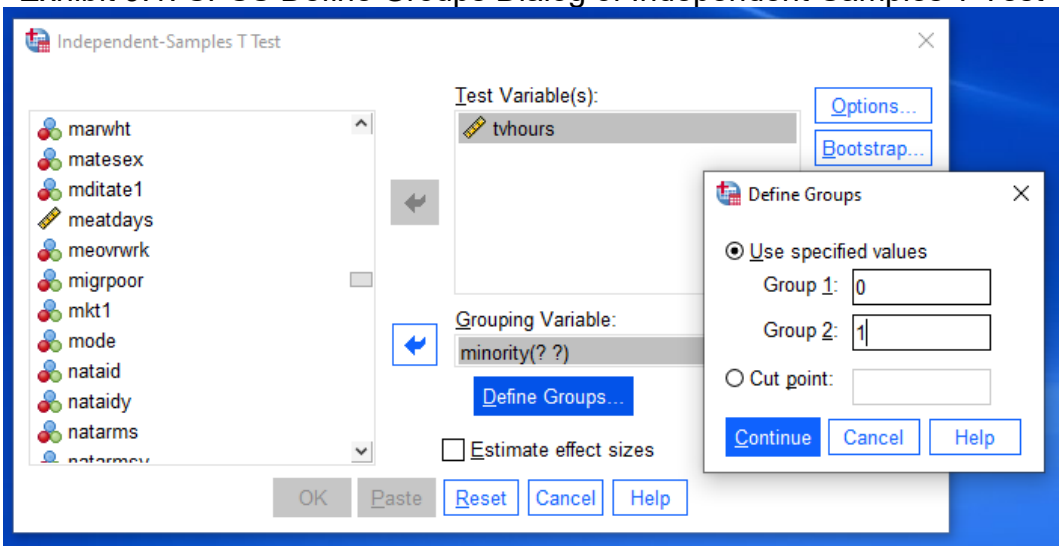


Exhibit 9.1 depicts the Define Groups dialog where we identify Group 1 as "0" (whites) and Group 2 as "1" (people of color). We need only click Continue and then OK to produce the *t* test output. Exhibit 9.2 provides a truncated version of the resulting SPSS output with the three crucial areas highlighted (the last four columns of the second table are not shown since we will not be using these).

Exhibit 9.2. Partial SPSS Output for *t* test of TVHOURS by MINORITY

Group Statistics

	minority people of color (dichotomous race)	N	Mean	Std. Deviation	Std. Error Mean
tvhours hours per day watching tv	0 White	2124	3.38	3.054	.066
	1 Person of Color	550	3.72	3.190	.136

Independent Samples Test

		Levene's Test for Equality of Variances				Significance	
		F	Sig.	t	df	One-Sided p	Two-Sided p
tvhours hours per day watching tv	Equal variances assumed	2.181	.140	-2.319	2672	.010	.020
	Equal variances not assumed			-2.260	828.448	.012	.024

In the Group Statistic box, we see that whites report watching an average of 3.38 hours of television in a typical day while people of color report watching an average of 3.72 hours of television in a typical day. The *t* test results will indicate whether the difference in these sample means is generalizable to the larger American adult populations of whites and people of color.

Notice the F and Sig. columns below the label, "Levene's Test for Equality of Variances." This is where we have to begin so that we know which row of *t* test results is accurate. At .140, the significance level of F is greater than our alpha of .054. Therefore, Levene's test is not statistically significant and equal variance can be assumed. So, we use the top "Equal variances assumed" row and see that the "Two-Sided p" value is less than our alpha of .054. Therefore, we reject the null of no difference between the means and conclude that the larger American adult populations of whites and people of color do differ in the amount of television that they watch. Since the mean value for people of color is higher, we find that on average, people of color watch significantly more television than whites.

The term "significantly" refers to statistical significance here. The *t* test is only a test of inference, so it is the sociologist's job to consider the practical importance of these mean differences. A difference of 0.34 hours per day is only 20 minutes. When attempting to interpret such findings, sociologists usually think about the "big picture" since we're looking at overall trends. Some whites certainly watch much more television than people of color and vice versa. The *t* test results simply focus on the average and take variation within each group into account.

Sociologists have documented abundant evidence surrounding institutional racism within contemporary American society. For instance, we know that people of color often do not have the same opportunities as whites in education and in the labor market. We also know a lot about racial residential segregation and the fact that communities of color often lack nearby access to parks and other outdoor recreational activities.[1]

Speaking of residences, let's see if homeownership has any impact upon television watching frequency in our second analysis. The DWELOWN variable is based upon a question on whether the respondent owns or rents their current dwelling (own = 1; rent = 2). Run the t test in SPSS and confirm that your output matches that of Exhibit 9.3.

Exhibit 9.3. Partial SPSS Output for *t* test of TVHOURS by DWELOWN

Group Statistics

	dwelown does r own or rent home?	N	Mean	Std. Deviation	Std. Error Mean
tvhours hours per day watching tv	1 own or is buying	1812	3.34	2.896	.068
	2 pays rent	784	3.71	3.555	.127

Independent Samples Test

		Levene's Test for Equality of Variances					
						Significance	
		F	Sig.	t	df	One-Sided p	Two-Sided p
tvhours hours per day watching tv	Equal variances assumed	31.530	<.001	-2.763	2594	.003	.006
	Equal variances not assumed			-2.549	1252.371	.005	.011

Is the Levene's test statistically significant? It is, so we must use the lower "Equal variances not assumed" row of the *t* test output. Is the *t* test statistically significant? Since the p value of .011 is less than our alpha of .054, we can generalize the difference in the mean scores to the larger populations. Those who report owning their homes watch an average of 3.34 hours of television in a typical day while those who rent watch an average of 3.71 hours. Thus, renters watch significantly more television than homeowners on average. Why do you think that this is the case? There are a number of plausible explanations. Those who rent typically have lower incomes, so they may not be able to afford as many entertainment options as those who own their homes.[2] Those who own may be more likely to live in larger households and be busier doing things with their families. Or, those who own may spend more time maintaining their homes and have less time for entertainment.

For our third analysis, let's test for sex differences in television viewing (TVHOURS by SEXBIRTH1). Is the *t* test statistically significant? It is not. So, the small difference in means (males = 3.38; females = 3.53) is not generalizable to the larger populations. Male and female respondents to the 2021 GSS do not statistically differ in their television viewing habits. Why not? The interpretation of the lack of group differences is usually the most difficult since we are explaining why X has no impact on Y. Sociologists of sex and gender often find that Americans have strong perceptions of sex differences and the various roles that men and women occupy. However, much empirical research demonstrates that males and

females are more similar than different.[3] Both, it turns out, watch too much TV!

Let's move on to our last examples. We have yet to explore the RELIG variable in the GSS. This question has been asked to respondents every year that the survey has been fielded since its inception in 1972: "What is your religious preference? Is it Protestant, Catholic, Jewish, some other religion, or no religion?" Exhibit 9.4 provides the 2021 frequency distribution of this nominal variable.

Exhibit 9.4. Frequency Distribution of RELIG

relig r's religious preference

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 protestant	1589	39.4	40.2	40.2
	2 catholic	824	20.4	20.9	61.1
	3 jewish	75	1.9	1.9	63.0
	4 none	1121	27.8	28.4	91.3
	5 other	55	1.4	1.4	92.7
	6 buddhism	47	1.2	1.2	93.9
	7 hinduism	30	.7	.8	94.7
	8 other eastern religions	2	.0	.1	94.7
	9 muslim/islam	25	.6	.6	95.4
	10 orthodox-christian	37	.9	.9	96.3
	11 christian	124	3.1	3.1	99.4
	12 native american	3	.1	.1	99.5
	13 inter-nondenominational	19	.5	.5	100.0
	Total	3951	98.0	100.0	
Missing	System	81	2.0		
Total		4032	100.0		

The institution of religion has undergone tremendous change over the past fifty years as secularism (the declining role of religion and increasing importance of "worldly" concerns) has grown.[4] For example, sociologists have learned that religious service attendance has been in decline. We also see notable shifts in preference and affiliation. The fastest growing

religious group in the United States is the "nones."[5] Notice that 28.4% of respondents to the 2021 GSS identify as having no religious preference. This has increased more than five-fold since 1972 when only 5.2% of respondents reported that they had no religious preference. Today, the American religious landscape is more complex and varied than in the past.

As you will recall, the independent samples *t* test allows us to test for mean differences on a scale variable between any two categories of another variable (and the level of measurement of X does not matter). The RELIG variable provides an opportunity to discuss a limitation of the *t* test and of random sample surveys in general. As we saw back in Chapter 5, the Central Limit Theorem underlies the concept of the sampling distribution and requires a sample size of at least 50 for inferential tests. The same logic is applied to *t* testing as the **sampling distribution of the difference between means** requires that categories of X (which can be thought of as the independent samples) have at least 50 cases to be used in *t* tests. As seen in Exhibit 9.4, seven of the 13 categories of RELIG contain fewer than 50 respondents. If it makes sociological sense, some of these categories could be combined through a recode command. For now, we are only going to focus on the first four response categories.

Back in Chapter 3, we explored the EDUC variable. This scale variable measures the years of formal education that the respondents have completed. Let's see if religious preference predicts years of education. Note that causality is not entirely clear here. One's religious preference is typically influenced by their parents and their upbringing. We also know that religiosity influences our attitudes and behaviors. Thus, religious preference could shape one's future career aspirations and the decision to attend college or not. Yet we also know that education influences our religiosity. Indeed, it is no coincidence that U.S. secularism has thrived as college attendance has seen tremendous growth. For some, the college experience can lead to greater confidence in science and greater concern for today's social problems. Religion has also become much more political and is cited for some policy and legal advocacy that many people can find short-sited or offensive.[6] While the direction of causality does appear to go in both directions, let's use RELIG as our X variable to test for an association with EDUC, our Y variable.

Do those who identify as Protestant differ from those who identify as Jewish in their educational attainment? To set this up in SPSS, we will

use the values of 1 (Protestant) and 3 (Jewish) in the Define Groups dialog. The output is provided as Exhibit 9.5.

Exhibit 9.5. Partial SPSS Output for *t* test of EDUC by RELIG (Protestant vs. Jewish Respondents)

Group Statistics

	relig r's religious preference	N	Mean	Std. Deviation	Std. Error Mean
educ highest year of school completed	1 protestant	1580	14.75	2.649	.067
	3 jewish	75	16.07	2.361	.273

Independent Samples Test

		Levene's Test for Equality of Variances					
						Significance	
		F	Sig.	t	df	One-Sided p	Two-Sided p
educ highest year of school completed	Equal variances assumed	4.952	.026	-4.220	1653	<.001	<.001
	Equal variances not assumed			-4.684	83.087	<.001	<.001

Hopefully, this is getting easy for you by now. Levene's test is statistically significant, so we cannot assume equal variances. The *t* test is highly statistically significant with a p value of <.001. Jewish respondents report an average of 16.07 years of education while Protestant respondents report an average of 14.75 years. Those who identify as Jewish have significantly higher educational attainment than those who identify as Protestant.

In our final example, let's compare Catholics (2) with those reporting that they have no religious preference ("none," category 4). Exhibit 9.6 provides the output. Here, Levene's test is not significant, so equal variances can be assumed. The *t* test is highly statistically significant with a p value of <.001. Respondents with no religious preference report an average of 14.95 years of education while Catholic respondents report an average of 14.47 years. Those with no religious preference have significantly higher educational attainment than those who identify as Catholic.

Exhibit 9.6. Partial SPSS Output for *t* test of EDUC by RELIG (Catholics vs. No Religious Preference)

Group Statistics

	relig r's religious preference	N	Mean	Std. Deviation	Std. Error Mean
educ highest year of school completed	2 catholic	821	14.47	2.966	.104
	4 none	1110	14.95	2.759	.083

Independent Samples Test

		Levene's Test for Equality of Variances					
						Significance	
		F	Sig.	t	df	One-Sided p	Two-Sided p
educ highest year of school completed	Equal variances assumed	.738	.390	-3.647	1929	<.001	<.001
	Equal variances not assumed			-3.608	1692.949	<.001	<.001

As we have seen, the independent samples *t* test is a powerful tool for the quantitative sociologist. This inferential test allows us to determine if a difference that is observed between two sample means can be generalized to the larger populations of interest. The major limitation of the test is that only one comparison between two groups is possible. This issue is resolved by employing the technique that is the topic of the next chapter, analysis of variance.

It is crucial to use the correct statistical test for the variables at hand. We also need to make sure that assumptions (such as minimum sample size requirements) are satisfied. Finally, quantitative sociologists must consider the practical importance of our findings. In our interpretations of why we found what we found, we must think about the extent to which it really matters. Statistically significant results can lack substantive meaning.

Key Terms

Independent samples *t* test, test variable, grouping variable, define groups, equality of variances, Levene's test, and sampling distribution of the difference between means

Endnotes

1. For example, see Chapter 15 in Wright, Erik Olin and Joel Rogers. 2024. *American Society: How It Really Works* Third Edition. New York, NY: W.W. Norton & Company.

2. Ginstein-Weiss, Michal, Trina R. Williams Shanks, Kim R. Manturuk, Clinton C. Key, Jong-Gyu Paik, and Johann K.P. Greeson. 2010. "Homeownership and Parenting Practices: Evidence from the Community Advantage Panel." *Children and Youth Service Review* 32: 774-782.

3. Ely, Robin and Irene Padavic. 2007. "A Feminist Analysis of Organizational Research on Sex Differences." *Academy of Management Review* 32 (4): 1121-1143.

4. Wuthnow, Robert. 2007*. After the Baby Boomers: How Twenty- and Thirty-Somethings Are Shaping the Future of American Religion*. Princeton, NJ: Princeton University Press.

5. Thiessen, Joel and Sarah Wilkins-LaFlamme. 2017. "Becoming a Religious None: Irreligious Socialization and Disaffiliation." *Journal for the Scientific Study of Religion* 56 (1): 64-82.

6. Audette, Andre P. and Christopher L. Weaver. 2016. "Filling Pews and Voting Booths: The Role of the Politicization in Congregational Growth." *Political Research Quarterly* 69 (2): 245-257.

This 2025 work is licensed under CC BY-NC-SA 4.0

Quantitative Sociology Essentials:
An Introduction to Survey Data Analysis using IBM® SPSS®

Ed Collom, Ph.D.
Professor of Sociology
California State University, Fullerton

Chapter 10. Analysis of Variance

In these final chapters of the text, we are learning inferential statistical tests to analyze associations involving dependent (Y) variables with a level of measurement of scale. In the previous chapter, we learned that *t* tests only permit comparison of the sample means from two groups. However, most categorical independent (X) variables of sociological interest have more than two categories. The **analysis of variance** (ANOVA) is our solution since it is based upon multiple group comparisons (see Exhibit 8.1).

Chapter Objectives

After reading this chapter, students should be able to:

- Determine with which variable types the analysis of variance is the appropriate inferential test statistic
- Produce an ANOVA with post hoc comparisons to test for an association between two variables
- Evaluate hypotheses by reporting generalizability and differences in group mean scores

ANOVA is a powerful, complex, and widely used statistical technique. The test comes in a variety of forms to accommodate many different research scenarios. We will only focus on what is referred to as the “one-way” ANOVA here. This is the simplest ANOVA and is often the choice of sociologists and survey researchers. Since this text aims to provide students with the essentials of how to analyze associations among all possible combinations of variable types, this chapter provides only a brief introduction to this complicated subject.

We can think of ANOVA as an extension of the *t* test for the difference in sample means. Instead of comparing just two means though, we are now able to compare multiple group means on a Y variable that is scale. The mathematical assumptions underlying ANOVA are somewhat different from those underlying *t*. On the positive side, there is no minimum sample size requirement for the groups we are comparing as there is in *t* testing. On the negative side, there is an assumption of homogeneity of variances. That is, the test assumes that the standard deviations on the Y variable for the various groups of X are similar. When they are not, more sophisticated analyses that go beyond the scope of this text are required. Statisticians have found that moderate violations of this assumption do not change the results. Moreover, with larger sample sizes, the ANOVA test is more robust and the homogeneity of variances matters less.[1] Given the objectives of this text and the fact that the 2021 General Social Survey is a very large sample, we will not be concerned about disparate standard deviations among the groups of the independent variable.

ANOVA is based upon the comparison of the variance on Y *between* the categories of the X variable to the amount of variance on Y *within* the categories of X. The computation can be thought of as comparing the differences among the various sample means from the groups of X to the various standard deviations of those groups. When the null hypothesis of no difference in the sample means is rejected, there is greater variation between the categories (the means) and less within (the standard deviations).

ANOVA's inferential test statistic is the ***F* ratio**. It is a global ("overall" or "omnibus") test determining whether there is some relationship between X and Y. As in the case of other inferential tests, the actual value of the *F* ratio does not have practical meaning to us. The only thing that concerns us is whether the p value or significance level of the test is less than or equal to our alpha of .054. If the *F* ratio is not statistically significant, we fail to reject the null and conclude that there is no association between the two variables that can be generalized.

<u>ANOVA in SPSS</u>

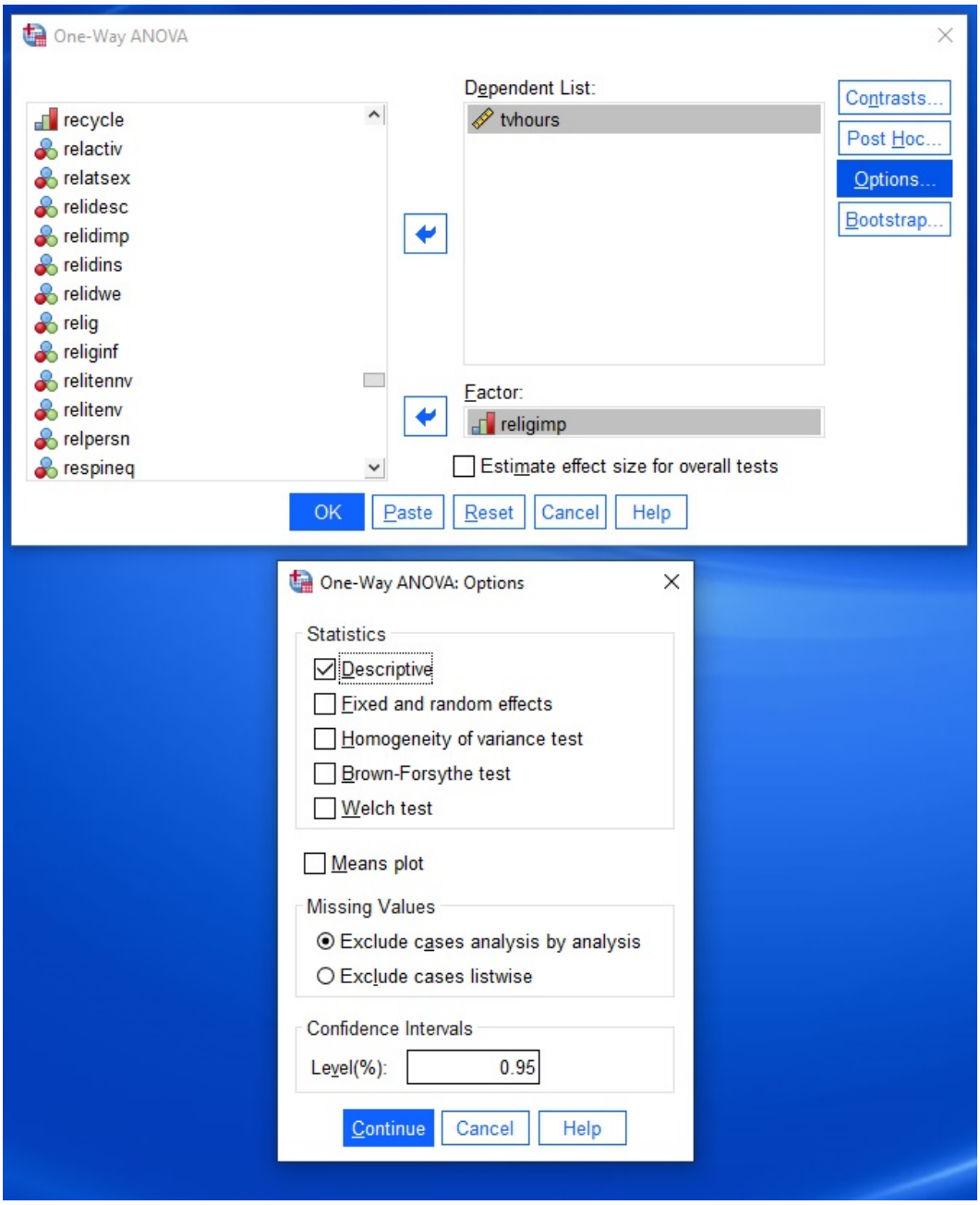
The path to the SPSS command is Analyze > Compare Means and Proportions > One-Way ANOVA. In the dialog, we must place the dependent (Y) variable that is measured at the scale level in the Dependent List box. The categorical independent (X) variable of interest

goes in the Factor field. We could then click OK and see the results of the *F* ratio. However, when the *F* ratio is statistically significant, we will need more information to determine which groups of the X variable have statistically different means from one another. Therefore, we will also always request "Descriptive" statistics from the Options dialog and "LSD" from the Post Hoc dialog. The descriptive statistics are needed so that we can see the mean scores on Y by the categories of X. **Post hoc tests** indicate which specific mean scores are statistically higher or lower than the other ones. In this text, we will employ the frequently used least significant difference (LSD) type.

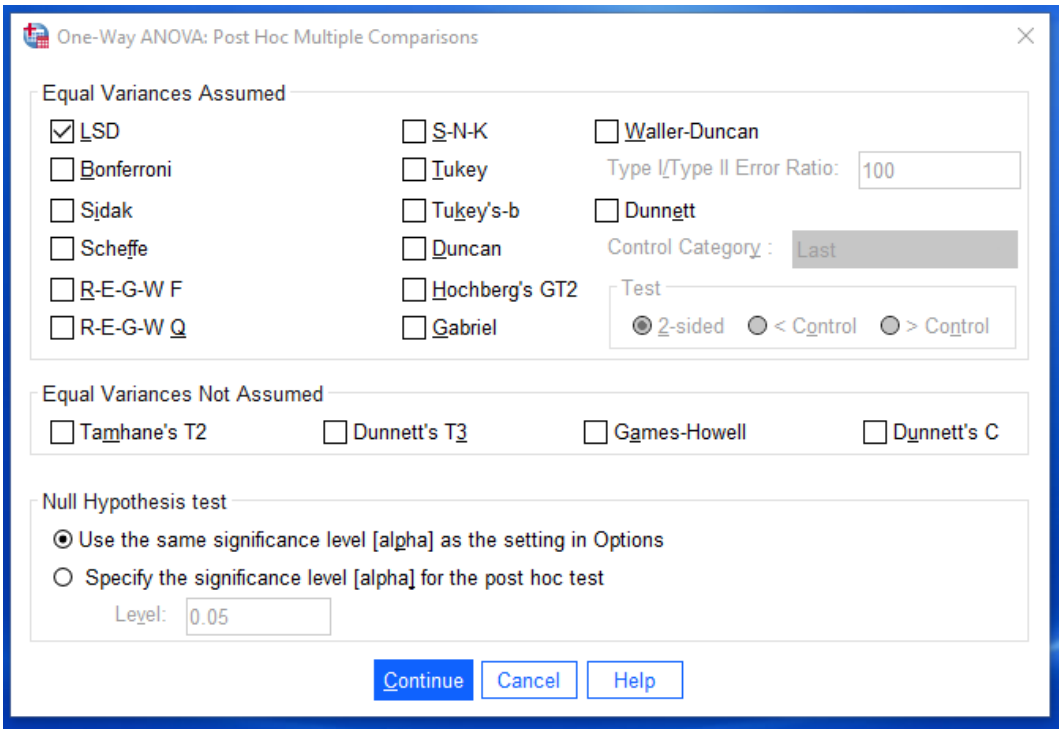
For our first example, we are going to continue to think about the influence that religion may have upon our behavior. The 2021 GSS contains a new variable, RELIGIMP: "How important is religion in your life – very important, somewhat important, not too important, or not at all important?" Let's determine if the importance of religion in one's life influences the amount of television one watches.

Exhibit 10.1 displays the completed dialogs in SPSS. Launch the GSS2021.sav data file and produce your first ANOVA of TVHOURS by RELIGIMP.

Exhibit 10.1. One-Way ANOVA Dialog in SPSS with Options and Post Hoc



(continued on next page)



As with most of its commands, the SPSS output contains more than what we need here. Exhibit 10.2 highlights the relevant portions of the ANOVA output. First, we should begin with the ANOVA table portion of the output in the middle. Is the *F* ratio statistically significant? Yes, we see that it is highly statistically significant and that we can reject the null of no differences among the means. In other words, the importance of religion in one's life does influence the amount of television one watches. The Descriptives at the top of the output contain the mean scores. Respondents who report that religion is "very important" in their lives watch an average of 3.56 hours of television in a typical day. Those for whom religion is "somewhat important" watch an average of 3.80 hours. The "not too important" group watches 3.51 hours. Respondents who report that religion is "not at all important" in their lives watch an average of 3.01 hours of television in a typical day.

Exhibit 10.2. SPSS ANOVA Output of TVHOURS by RELIGIMP

Descriptives

tvhours hours per day watching tv

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1 very important	746	3.56	3.473	.127	3.31	3.81	0	24
2 somewhat important	624	3.80	3.013	.121	3.56	4.03	0	24
3 not too important	436	3.51	3.253	.156	3.20	3.82	0	24
4 not at all important	598	3.01	2.615	.107	2.80	3.22	0	20
Total	2404	3.48	3.129	.064	3.35	3.60	0	24

ANOVA

tvhours hours per day watching tv

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	199.593	3	66.531	6.845	<.001
Within Groups	23326.103	2400	9.719		
Total	23525.696	2403			

Post Hoc Tests

Multiple Comparisons

Dependent Variable: tvhours hours per day watching tv

LSD

(I) religimp how important is religion	(J) religimp how important is religion	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1 very important	2 somewhat important	-.236	.169	.162	-.57	.10
	3 not too important	.052	.188	.780	-.32	.42
	4 not at all important	.550*	.171	.001	.21	.89
2 somewhat important	1 very important	.236	.169	.162	-.10	.57
	3 not too important	.289	.195	.138	-.09	.67
	4 not at all important	.786*	.178	<.001	.44	1.14
3 not too important	1 very important	-.052	.188	.780	-.42	.32
	2 somewhat important	-.289	.195	.138	-.67	.09
	4 not at all important	.497*	.196	.011	.11	.88
4 not at all important	1 very important	-.550*	.171	.001	-.89	-.21
	2 somewhat important	-.786*	.178	<.001	-1.14	-.44
	3 not too important	-.497*	.196	.011	-.88	-.11

*. The mean difference is significant at the 0.05 level.

The last portion of the output is the most confusing, but crucial in an ANOVA. The Multiple Comparisons matrix reports the post hoc least significant difference test results. This output identifies which groups statistically differ from one another in their television viewing. The matrix reports one value/response category of X and compares it to the other three for statistical differences. The Sig. column lists the p value for each

comparison. We are looking for Sig. values that are less than or equal to .054 to conclude that the groups differ.

The group 1 (very important) and group 2 (somewhat important) means do not statistically differ since Sig. = .162. In other words, those who find religion “very important” or “somewhat important” watch similar amounts of TV. Group 1 and group 3 also do not differ (Sig. = .780). The difference in mean scores between group 1 and group 4 is statistically significant (Sig. = 001). Now, go back up to the Descriptives output to see which group is higher and which is lower. Respondents who report that religion is “very important” in their lives watch significantly more television than those who find religion “not at all important.”

The next portion of the matrix compares group 2 (somewhat important) to each of the other groups. As you can see, there is redundant information as each comparison is listed twice in this matrix. The group 1 to group 2 comparison in the first part is the same comparison as group 2 to group 1 in the second part. The Sig. values will be identical by definition (.162 here). The only group that has a statistically different mean from the “somewhat important” respondents is group 4 (not at all important). Looking at the mean scores in the Descriptives section again, we see that respondents who report that religion is “somewhat important” in their lives watch significantly more television than those who find religion “not at all important.”

The third part of the Multiple Comparisons matrix focuses on group 3 (not too important). Here, there is only one statistical difference (with group 4). Respondents who report that religion is “not too important” in their lives watch significantly more television than those who find religion “not at all important.”

The final portion of the matrix is for group 4 (not at all important). This information is entirely redundant since we have already seen all three of these comparisons. However, as it turns out, the trend is very clear in this section since group 4 differs from all of the other groups. And, as we saw in the earlier listings, groups 1, 2, and 3 do not differ from one another. So, we can summarize the findings in one sentence: Respondents who report that religion is “not at all important” in their lives watch significantly less television than all of the other groups.

As you can see, the Multiple Comparisons matrix can be confusing. You may find it easiest to write down the statistically significant mean differences. The following method (using the values of the response categories of X) is one way to do it – 1:4; 2:4; and 3:4. This tells us that group 4 is different from the other three and groups 1, 2, and 3 do not differ from one another.

Our X variable, RELIGIMP, has four categories. When you have four mean scores, there are six possible mean difference comparisons. The formula for computing the number of comparisons possible uses the term k which represents the number of categories in a variable: k(k-1)/2. 4x3=12/2=6. Using this formula, we find that there are 10 comparisons for an X variable with 5 categories, 15 for an X with 6 categories, 21 for an X with 7 categories, 28 for an X with 8 categories, and so forth. In other words, ANOVAs involving X variables with many categories can be very complicated.

So, why is it that those who report that religion is "not at all important" in their lives watch the least amount of TV? There are many possible explanations. Recall from Chapter 1 that the second step of the research process is reviewing the literature. Existing literature such as sociological theory or previous empirical research findings typically guide our research hypotheses as well as our interpretation of the results. However, earlier studies found either no association or that religious people actually watch television less frequently.[2] If our results are not supported by any literature, then we can make an educated guess. We know that religion and media are both major social institutions. Those who find religion as "not at all important" are independent from and perhaps even reject the social institution of religion. Thus, it is plausible that these same people will reject mainstream media and television as well. We could also consider the fact that many Americans are extremely busy and have limited down time. Some people work loads of hours every week and simply may not have the time to be concerned with religion nor to watch television.

Let's move on. We have used the DEGREE variable numerous times throughout this text. Does educational attainment influence the amount of television one watches? Run an ANOVA of TVHOURS by DEGREE and confirm that your output matches that of Exhibit 10.3.

Exhibit 10.3. Partial SPSS ANOVA Output of TVHOURS by DEGREE

tvhours hours per day watching tv

	N	Mean
0 less than high school	161	4.17
1 high school	1066	3.99
2 associate/junior college	252	3.59
3 bachelor's	672	2.89
4 graduate	517	2.84
Total	2668	3.46

ANOVA

tvhours hours per day watching tv

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	794.964	4	198.741	21.121	<.001
Within Groups	25058.363	2663	9.410		
Total	25853.326	2667			

Multiple Comparisons

Dependent Variable: tvhours hours per day watching tv

LSD

(I) degree r's highest degree	(J) degree r's highest degree	Mean Difference (I-J)	Std. Error	Sig.
0 less than high school	1 high school	.181	.259	.486
	2 associate/junior college	.576	.309	.063
	3 bachelor's	1.273*	.269	<.001
	4 graduate	1.328*	.277	<.001
1 high school	0 less than high school	-.181	.259	.486
	2 associate/junior college	.396	.215	.066
	3 bachelor's	1.093*	.151	<.001
	4 graduate	1.147*	.164	<.001
2 associate/junior college	0 less than high school	-.576	.309	.063
	1 high school	-.396	.215	.066
	3 bachelor's	.697*	.227	.002
	4 graduate	.752*	.236	.001
3 bachelor's	0 less than high school	-1.273*	.269	<.001
	1 high school	-1.093*	.151	<.001
	2 associate/junior college	-.697*	.227	.002
	4 graduate	.055	.179	.760
4 graduate	0 less than high school	-1.328*	.277	<.001
	1 high school	-1.147*	.164	<.001
	2 associate/junior college	-.752*	.236	.001
	3 bachelor's	-.055	.179	.760

Is the *F* ratio statistically significant? Yes, we see that it is highly statistically significant and that we can reject the null of no differences among the means. In other words, educational attainment does influence the amount of television one watches. While inspecting the mean scores, notice that there is a downward trend. Those who did not graduate from high school watch an average of 4.17 hours per day, those with high school degrees watch an average of 3.99 hours, those with junior college degrees watch an average of 3.59 hours, those with Bachelor's degrees watch an average of 2.89 hours, and those with graduate school degrees watch an average of 2.84 hours of television in a typical day.

Now, let's report the findings from the Multiple Comparisons matrix. Remember that we are searching for comparisons in which Sig. ≤ .054. Using the method discussed in the previous example, here are the groups that statistically differ from one another – 0:3; 0:4; 1:3; 1:4; 2:3; and 2:4. Notice that groups 0, 1, and 2 each statistically differ from groups 3 and 4, but are not different from one another. These findings can be summarized rather succinctly: Those who have earned their Bachelor's or graduate school degrees watch significantly fewer hours of television in a typical day than those with lower educational attainment. Why? Perhaps it's because those with more education tend to have better remunerating jobs. They may have greater ability to afford other forms of entertainment. They are also likely to work more hours, so they may just lack the time to watch as much TV as the others.

Let's move on to our third example. Does the frequency with which one reads a newspaper influence the amount of television one watches? The NEWS variable is an ordinal one ranging from "every day" (1) to "never" (5). Produce the ANOVA. Is the *F* ratio statistically significant? It is not since Sig. = .365. Therefore, we must conclude that the frequency of reading newspapers is not associated with the frequency of watching TV. These forms of media consumption are independent of one another.

Now, let's return to RELIGIMP as our X variable and consider its impact upon weekly hours of internet usage (WWWHR). Run an ANOVA of WWWHR by RELIGIMP and confirm that your output matches that of Exhibit 10.4.

Exhibit 10.4. Partial SPSS ANOVA Output of WWWHR by RELIGIMP

wwwhr www hours per week

	N	Mean
1 very important	680	12.57
2 somewhat important	585	12.98
3 not too important	419	16.93
4 not at all important	580	18.96
Total	2264	15.12

ANOVA

wwwhr www hours per week

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	17019.159	3	5673.053	18.838	<.001
Within Groups	680586.163	2260	301.144		
Total	697605.322	2263			

(I) religimp how important is religion	(J) religimp how important is religion	Mean Difference (I-J)	Std. Error	Sig.
1 very important	2 somewhat important	-.407	.979	.677
	3 not too important	-4.356*	1.078	<.001
	4 not at all important	-6.388*	.981	<.001
2 somewhat important	1 very important	.407	.979	.677
	3 not too important	-3.949*	1.111	<.001
	4 not at all important	-5.981*	1.017	<.001
3 not too important	1 very important	4.356*	1.078	<.001
	2 somewhat important	3.949*	1.111	<.001
	4 not at all important	-2.032	1.113	.068
4 not at all important	1 very important	6.388*	.981	<.001
	2 somewhat important	5.981*	1.017	<.001
	3 not too important	2.032	1.113	.068

Is the *F* ratio statistically significant? Yes, it is highly statistically significant and we can reject the null of no differences among the means. The importance of religion in one's life does influence the frequency with which one uses the internet. Notice the upward trend among the mean scores. The groups that statistically differ from one another are 1:3; 1:4; 2:3; and 2:4. Groups 1 and 2 are similar and Groups 3 and 4 are similar. They do not differ from each other, but they do differ from the other block. Respondents who report that religion is "very important" or "somewhat important" use the internet significantly less often than those who find religion "not too important" or "not at all important." Perhaps those for whom religion is important in their lives are more likely to spend their free time acting on its behalf (attending services, reading texts, etc.).[3]

The final example that we will explore in this chapter tests whether one's political views influence what they believe is the ideal number of children that families should have. We've already seen the political views variable in earlier chapters. However, for this analysis, we will use the simplified POLVIEW3 version (liberal=1; moderate=2; conservative=3). The KIDSIDEAL variable is based on the question, "What do you think is the ideal number of children for a family to have?"

You should find that the *F* ratio of this ANOVA is significant. Those who identify as liberal have an average ideal number of children of 2.26, moderates are at 2.46, and conservatives have an average of 2.54. The Multiple Comparisons post hoc results indicate that on average, those who identify as liberal have a lower ideal number of children than those identifying as moderate or conservative. Why do liberals prefer that families have fewer kids? Liberals may be less attached to the institution of family and the idea of the "American Dream" (buying a house in the suburbs and having children). It could also be that liberals are more likely to be environmentalists who are concerned about the "carrying capacity" of Earth and believe that smaller families are more sustainable.

As evident in this chapter, ANOVA is a powerful and complex statistical tool. Sociologists and survey researchers regularly employ it with scale dependent variables since many of our independent variables are categorical ones that are not dichotomous. This inferential test allows us to determine if differences observed among multiple sample means can be generalized to the larger population.

We have almost exhausted all of the possible bivariate combinations of the three variable types: nominal, ordinal, and scale. In the final chapter of this text, we will learn how to perform regression analyses involving two scale variables. Regression techniques are the most widely used ones in sociology and easily allow us to estimate multivariate models with multiple independent variables.

Key Terms

Analysis of variance (ANOVA), *F* ratio, and post hoc tests

Endnotes

1. Ramachandran, Kandethody M. and Chris P. Tsokos. 2020. *Mathematical Statistics with Applications in R* 3ed. Cambridge, MA: Academic Press.

2. For the former, see McClure, Paul K. 2020. “The Buffered, Technological Self: Finding Associations between Internet Use and Religiosity.” *Social Compass* 67 (3): 461-478. For the latter, see Bobkowski, Piotr S. 2009. “Adolescent Religiosity and Selective Exposure to Television.” *Journal of Media and Religion* 8 (1): 55-70.

3. See Armfield, Greg G. and R. Lance Holbert. 2003. “The Relationship Between Religiosity and Internet Use.” *Journal of Media and Religion* 2 (3): 129-144 and Bobkowski, Piotr S. 2009. “Adolescent Religiosity and Selective Exposure to Television.” *Journal of Media and Religion* 8 (1): 55-70.

This 2025 work is licensed under CC BY-NC-SA 4.0

Quantitative Sociology Essentials:
An Introduction to Survey Data Analysis using IBM® SPSS®

Ed Collom, Ph.D.
Professor of Sociology
California State University, Fullerton

Chapter 11. Regression and Correlation

The final topic on performing bivariate analyses of survey data concerns associations among two scale variables. Regression techniques are more widely used by quantitative sociologists than any of the other inferential statistical tests covered in this text. This chapter covers scattergrams, linear regression, Pearson's r correlation coefficient, and provides an introduction to estimating the impact of several independent (X) variables upon a dependent variable in multiple linear regression models.

As we learned earlier, scale variables have too many categories to use them in crosstabulations. If our independent variable is nominal or ordinal, then we would use means testing (*t* tests or ANOVA). Regression is a powerful tool to test for the association between two scale variables.

Chapter Objectives

After reading this chapter, students should be able to:

- Determine with which variable types regression techniques are the appropriate inferential tests
- Produce and report scattergram results to visualize the association between two scale variables and assess direction and strength
- Produce and report the results from bivariate linear regressions and correlation matrices to test for associations between two scale variables
- Produce and report the results from a multiple linear regression model to test for the impact of multiple independent variables upon a dependent variable and assess the strength of each predictor

Scattergrams

Scattergrams, also known as scatterplots, can be thought of as a visual crosstabulation. They plot two scale variables on a graph and enable us to see if responses to the X variable influence responses to the Y variable. When requesting a scattergram from SPSS, we always place the independent (X) variable on the horizontal (X) axis and the dependent (Y) variable on the vertical (Y) axis. The cases (respondents) are depicted as dots on the graph. Our visual analysis will allow us to assess if an association exists, and if so, the direction and relative strength of the relationship.

A key requirement of regression techniques is that the association between the two scale variables is a linear one. Scattergrams allow us to determine if the **linearity assumption** is met. A linear relationship is one in which a straight line summarizes the relationship relatively well. The **regression line** is the best-fitting straight line depicting the association between the two variables. It is an estimate of the association as a whole and will always have error since we do not find perfect associations when analyzing survey data. The distance on the graph between each individual case and the regression line is the **residual**. It is the difference between the observed values on X and Y and those predicted by the regression line. When SPSS fits the straight line to the data, it does so by minimizing the error in the model.

As we did in Chapter 8 on crosstabulations, let's begin with sample scattergrams illustrating the extreme possibilities. Suppose that there is a scale X variable and a scale Y variable with values ranging from 0 to 10. Exhibit 11.1.a. depicts the perfect positive association. The eleven dots on the graph are **case configurations**, where respondents fall on X and on Y. The dots can represent more than one respondent or case. Here, we see that those with low values on X, have low values on Y. Those who are moderate on X are also moderate on Y. Those with high values on X, have high values on Y. The regression line has a positive slope and perfectly intersects the cases. There is no residual. With any given value of X, we could predict one's outcome on Y.

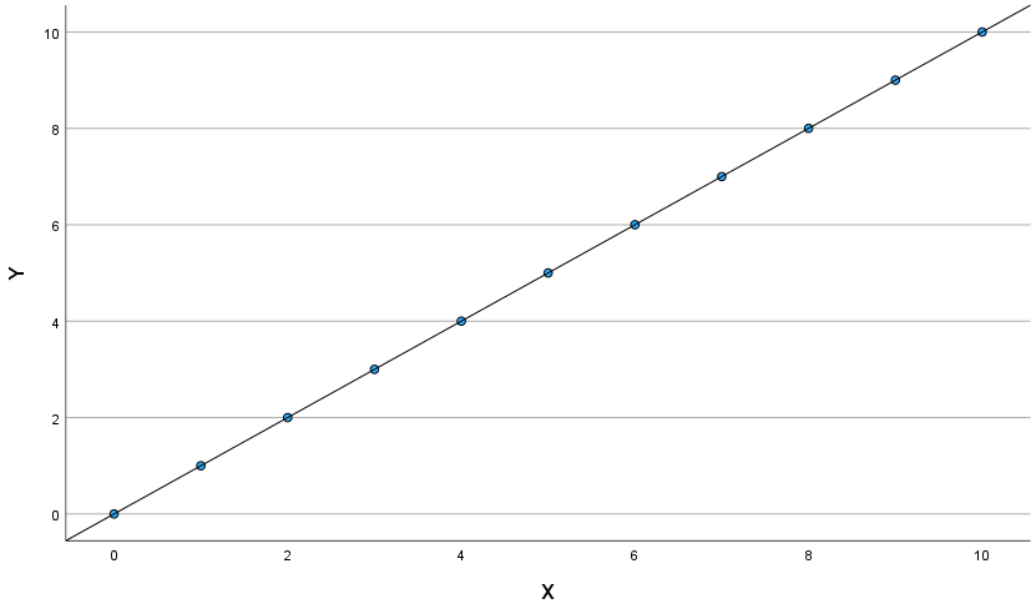
Exhibit 11.1.b. illustrates the perfect negative association. The case configuration is still perfect, but in the opposite direction. Low values of X correspond with high values on Y. Those with high values on X, have low values on Y. The slope of the regression line is negative and there is still no residual. As respondents increase on X, they decrease on Y.

Exhibit 11.1.c. depicts the perfect nonassociation. That is, there is absolutely no relationship between X and Y here. The cases are scattered across the entire field of the graph. There is no pattern. Those who are low on X, can be low, moderate, or high on Y. The regression line is perfectly parallel to the X axis. The slope of this line is neither positive nor negative. It is zero and indicates no association.

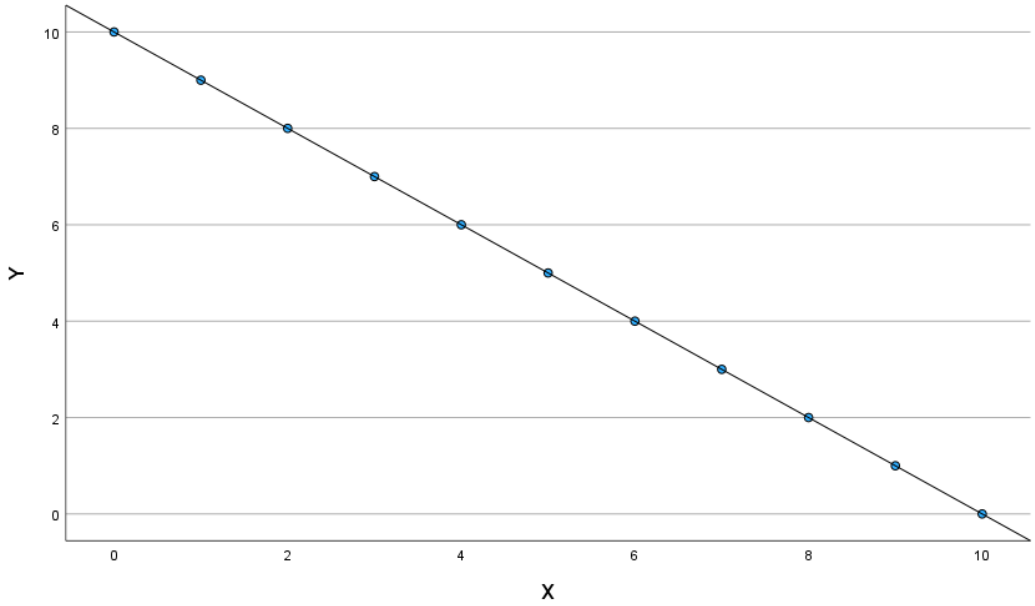
These three samples are the extreme possibilities of scattergrams. They are ideal types in that they do not exist in actual sociological survey research. While we do regularly find nonassociations, it is very rare that they are perfect nonassociations in which the cases are perfectly scattered all across the graph.

Exhibit 11.1. Scattergram Examples

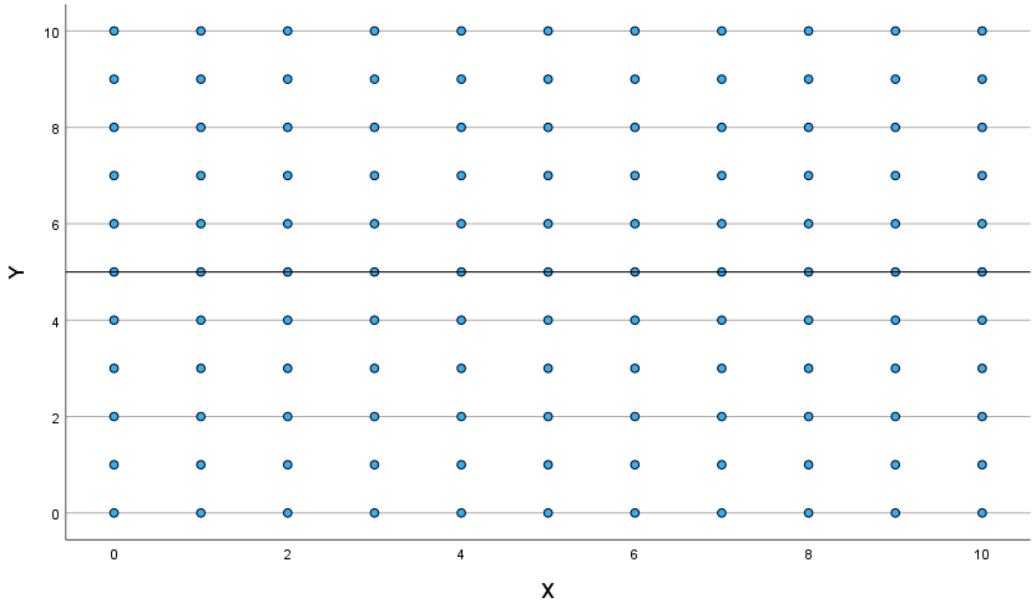
a. Perfect Positive Association



b. Perfect Negative Association



c. Perfect Nonassociation



Since scattergrams plot every case, they can be very confusing and busy-looking with large datasets such as the 2021 General Social Survey. Therefore, we are going to create scattergrams using a simpler dataset.

The “states.sav” datafile has only 51 cases. Instead of individuals, the unit of analysis is the 50 U.S. states and the District of Columbia. In addition to the state name, there are 12 variables measuring different aspects of those residing within each state.[1]

The SPSS command is Graphs > Scatter/Dot. The Scatter/Dot dialog offers five options, but we will only use the default Simple Scatter here, so you can just click Define. For our first example, let’s consider if median age (MEDIANAGE) impacts the labor force participation rate (LFPARTIC) within a state. Median age identifies the midpoint of the age distribution within each state and allows us to distinguish states with more younger residents versus states with more older residents. Labor force participation rate is the percentage of all residents of working age who are employed or actively seeking work.

Exhibit 11.2. Simple Scatterplot Dialog in SPSS

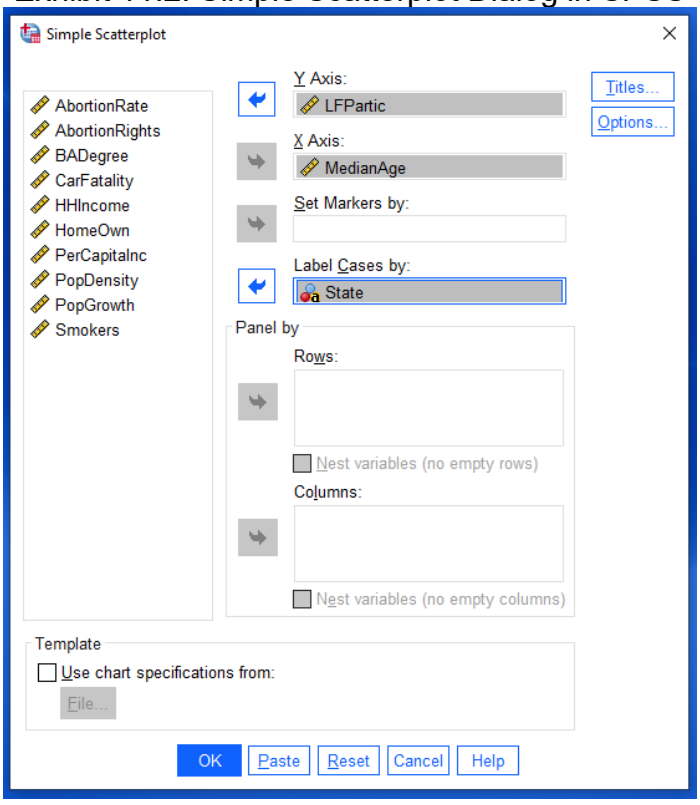


Exhibit 11.2 shows the Simple Scatterplot dialog in SPSS to produce the scattergram. MEDIANAGE is the independent variable here, so it goes on the X axis. LFPARTIC is the dependent variable, so it goes on the Y axis. By placing the STATE variable in the Label Cases by box, we will be able to use labels and see where each of the cases fall on X and Y. Click OK.

Exhibit 11.3. Scattergram of LFPARTIC by MEDIANAGE

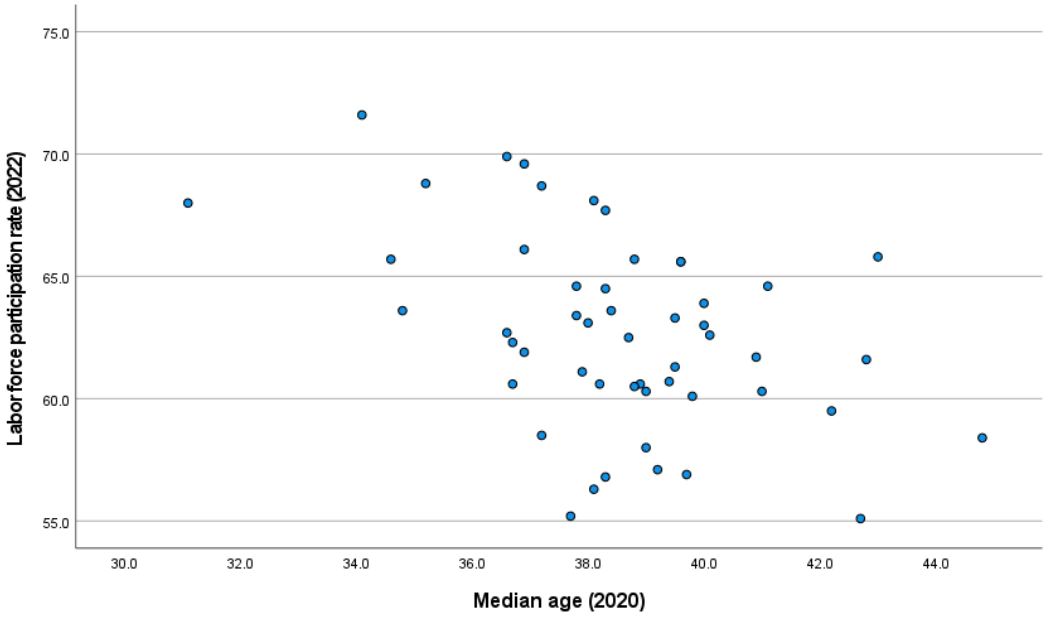
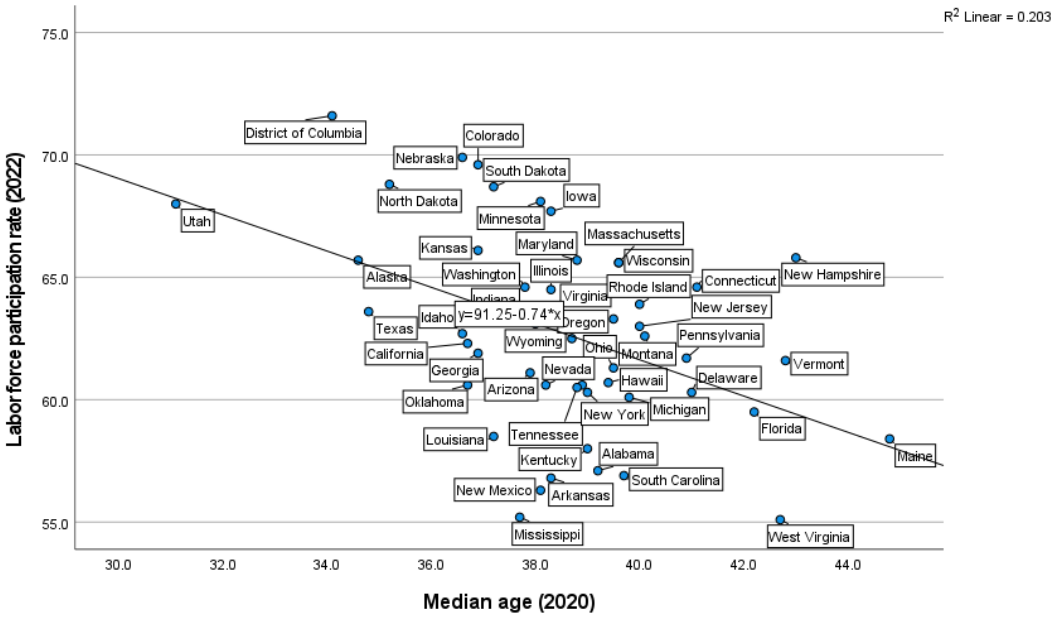


Exhibit 11.3 illustrates the resulting scattergram. Notice the value scale of the X axis. According to descriptive statistics from this dataset, the median age ranges from 31.1 to 44.8 years. Going from left to right, the median age of the residents within each state increases. On the Y axis, the labor force participation rate ranges from 55.1% to 71.6%. As you move up, the labor force participation rate increases within each state. Would a straight line summarize this association? If so, would the slope of the line be positive or negative? It appears that a line with a negative slope would capture the overall trend.

Now, we need to open the Chart Editor. Double-click the graph or right-click over it and select Edit from the dropdown menu. On the right side of the second bar at the top of the Chart Editor there are a series of icons. Click the one that looks like a scattergram with a line through it (Add Fit

Line at Total) to add the regression line ( ). The default line is what we want, so you can Close the Properties dialog when that pops up. Then, click on any of the blue dots (cases) to identify that you want to further edit the chart. Next, click the Show Data Labels icon that resembles a bar chart ( ). Once the state names appear, click Close on the Properties pop up and then click the X in the top right corner of the Chart Editor to close it and return to the SPSS output. The revised scattergram should appear as seen in Exhibit 11.4.

Exhibit 11.4. Scattergram of LFPARTIC by MEDIANAGE
with Regression Line and Data Labels



The slope of the regression line is indeed negative. As the median age within the states increases, the labor force participation rate decreases. This is sensible as most people age out and retire from the labor force at some point. Some of the states are on or very close to the regression line estimating the association between X and Y. That is, the labor force participation rate in states such as Utah, Alaska, Wyoming, Oregon, Ohio, Delaware, Florida, and Maine is very well predicted by the median age. In some states, the residual is quite large and their case configuration deviates from the overall trend. For example, Mississippi and New Hampshire are outliers. The former has a much lower labor force

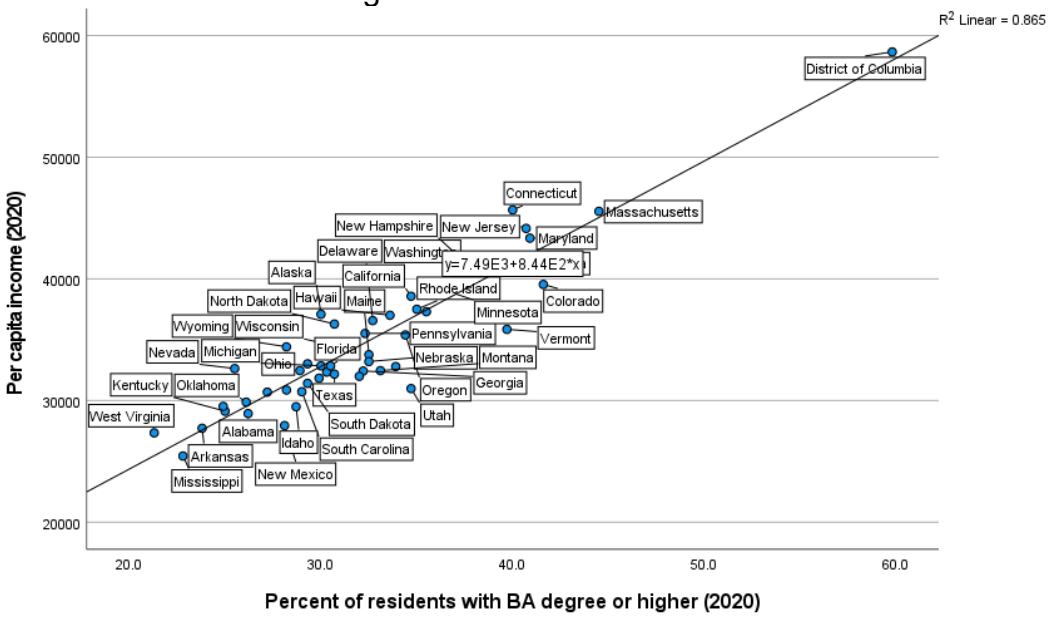
participation rate and the latter a higher one than is predicted by the regression line.

The slope of the regression line immediately indicates the direction of the association. There are two indicators of the strength of the association between two variables in a scattergram. First, the steeper the slope, the stronger the association. The flatter the slope and the closer it is to paralleling the X axis, the weaker the association. The second indicator of strength surrounds the **clustering** of the cases. The tighter the clustering of the cases around the regression line, the stronger the association. As we shall see soon, we can complement scattergrams with the appropriate measure of association for scale variables to achieve a precise assessment. You likely have already noticed that the regression line formula appears in the middle of the line and an additional statistic, R-square, is reported in the top right corner (more on these later).

Does the percent of residents with a college degree (BADEGREE) influence the per capita income (PERCAPINC) within each state? The X variable is the percent of the population within each state that has earned a Bachelor's degree or higher. Per capita income is a statistic frequently used in economics and sociology.[2] It is the sum of all of the income earned by people in an area divided by the total number of residents. Produce the scattergram in SPSS and confirm that it matches that in Exhibit 11.5.

As you can see, the direction to the association is clearly positive. The greater the percentage of residents with college degrees, the higher the per capita income within a state. Sociologists know that those with degrees tend to have better remunerating jobs. People with degrees are also more likely to relocate to areas with better paying jobs.[3] Thus, it is no surprise that states with more educated people have higher per capita incomes. Notice that the clustering of the cases near the regression line is much tighter in this example. In other words, this association is stronger than the previous one. It is interesting to see that D.C. is an outlier, but is still nearly perfectly predicted by the estimate. Residents of D.C. are highly educated and highly remunerated.

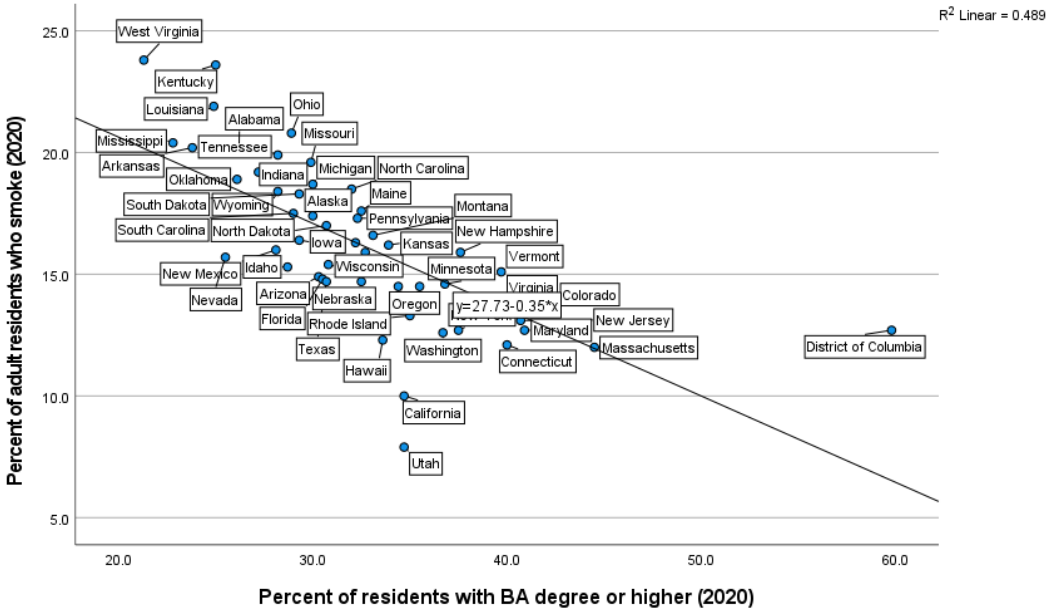
Exhibit 11.5. Scattergram of PERCAPINC by BADEGREE
with Regression Line and Data Labels



Let's produce one more scattergram. Does the percent of residents with a college degree (BADEGREE) influence the percent of adult residents who smoke (SMOKERS) within each state? Exhibit 11.6 presents the results.

The value scale of the Y axis shows the variance in smoking rates. You can look up the exact values per state in the Data View of the SPSS Data Editor. Smoking rates are lowest in Utah (only 7.9% of adult residents) and highest in West Virginia (23.8%). Residents of West Virginia are three times more likely to smoke than those of Utah. Sociologists understand the power of social networks and contagion of both positive and negative behaviors.[4] California has the second lowest smoking rate in the nation. The less that we are surrounded by people who smoke, the less likely it is that we will smoke. Of course, state legislation such as cigarette taxes also discourage this unhealthy habit. This scattergram clearly depicts a negative association. The clustering of the cases is moderate (lower than the previous example, but higher than the first one).

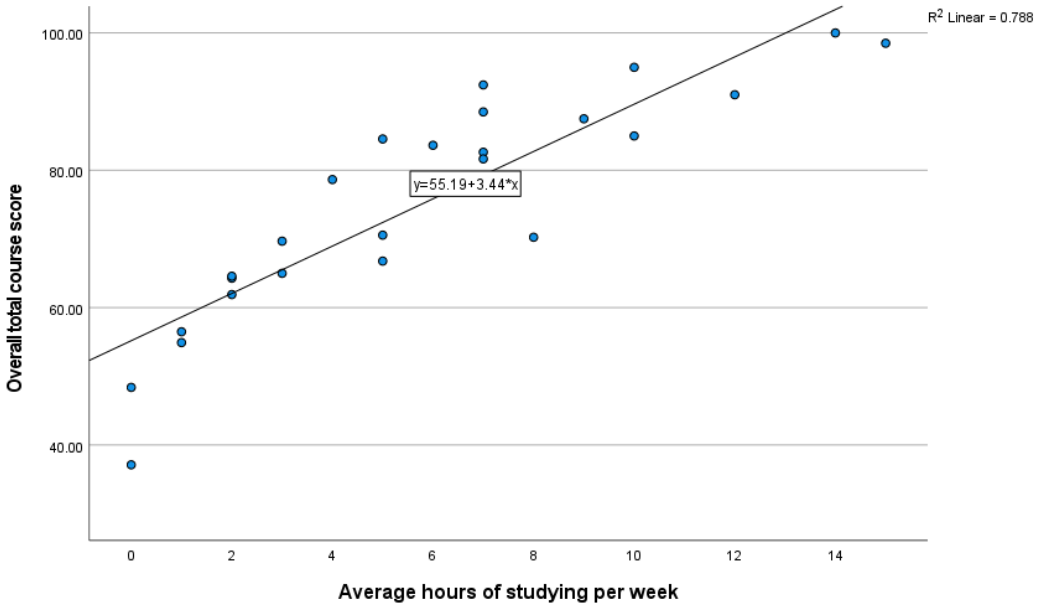
Exhibit 11.6. Scattergram of SMOKERS by BADEGREE
with Regression Line and Data Labels



Now, let's learn about that formula we keep seeing in the middle of the regression line. We know that the regression line is the single, best-fitting straight line (with the least amount of error) that summarizes the relationship between the two variables. We will use a fictitious dataset of 25 students for this example (SampleClassData.sav). Suppose that at the end of the semester, a professor asked students to report the average number of hours that they studied for the course each week (STUDYHOURS). Then, the professor is able to see if that X variable is associated with each student's overall total course score on a 100% scale (TOTALSCORE). Exhibit 11.7 presents this fictitious scattergram.

As we can see, the association is clearly positive and the case clustering around the regression line is relatively tight. Let's dissect that formula in the graph: Y = 55.19 + 3.44(X). The **regression line formula** is relatively simple as it contains only four terms. We should immediately understand two of them. Y is the predicted score on the dependent variable. That is, the regression line formula results in the value of Y. To compute it, information about X, the score on the independent variable, is required.

Exhibit 11.7. Fictitious Scattergram of TOTALSCORE by STUDYHOURS



The two reported numeric values in the equation above are for the other two components, a and b. The "a" in the formula is the **Y-intercept**. This is the point where the regression line crosses the Y axis. This is also the predicted value of Y in cases in which X is zero. If X = 0, then the product of b times X will automatically be zero and leave only the value of a left in the equation. The "b" in the formula is the slope of the regression line. This value tells us the change in the value of Y for each one unit change in the value of X. As we already know, the slope can be positive or negative. So, with a positive slope, the product of b times X is added to a, the Y-intercept. With a negative slope, the product of b times X is subtracted from a. Exhibit 11.8 summarizes the formula.

Exhibit 11.8. Regression Line Formula

Y = a + b(X)

Y = the predicted score on the dependent variable
X = the score on the independent variable
a = the Y-intercept (point where line crosses Y axis)
b = slope of regression line (change in Y with unit change in X)

Going back to Exhibit 11.7, we see that the formula for the fictitious scattergram is reported as Y = 55.19 + 3.44(X). This equation can be used to obtain a predicted value of the total course score given a student's average hours of studying each week. For example, a student who did not study at all (0 hours per week), would be expected to have a final course score of 55.19% (since 3.44 times 0 is 0). A student who studies for 10 hours per week is predicted to have a total course score of 89.59% [Y = 55.19 + 3.44(10) = 89.59]. The value of the slope b, +3.44, indicates that on average, each additional hour of study time per week results in a 3.44 percentage point increase in one's total course score. This simple, yet powerful, formula is the basis of linear regression.

Linear Regression

While scattergrams visualize the association between two scale variables, **linear regression** is the mathematical model summarizing the impact of an X variable on a Y variable. The formulas for the calculation of "a" and "b" go beyond the scope of this text and get rather complicated (involving the mean, deviations from the mean, and measures of dispersion). Let's return to the 2021 GSS to produce our first linear regression in SPSS. In the previous chapter, we tested the influence of DEGREE upon TVHOURS using ANOVA. Now, let's use the scale variable EDUC (years of formal education) as the predictor of TVHOURS in a linear regression.

The path to the SPSS command for linear regression is Analyze > Regression > Linear. The dialog is very simple as we need only move TVHOURS into the Dependent box and EDUC into the Independent(s) box. No other options are required as the default output suffices. Exhibit 11.9 presents the output with the crucial areas highlighted. There are six that we need to focus upon.

First, begin with the Coefficients table at the bottom. The value of the Y-intercept (a) is found in the Constant row under the "B" column (Unstandardized Coefficients). It is reported as 5.898. Just by knowing this value, we are able to determine that those with no formal education (0 years) watch an average of 5.898 hours of television in a typical day.

Exhibit 11.9. SPSS Linear Regression Output of TVHOURS on EDUC

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.150[a]	.022	.022	3.081

a. Predictors: (Constant), educ highest year of school completed

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	575.777	1	575.777	60.654	<.001[b]
	Residual	25117.899	2646	9.493		
	Total	25693.676	2647			

a. Dependent Variable: tvhours hours per day watching tv

b. Predictors: (Constant), educ highest year of school completed

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.898	.319		18.502	<.001
	educ highest year of school completed	-.165	.021	-.150	-7.788	<.001

a. Dependent Variable: tvhours hours per day watching tv

Next, we need to determine if the independent variable EDUC is a statistically significant predictor of the dependent variable TVHOURS. Looking across the EDUC row (last in the table), the last column is labeled Sig. and provides the level of statistical significance. It is reported as <.001. This is far below our alpha of .054, so we can conclude that EDUC is a statistically significant predictor of TVHOURS.

The slope of the regression line (b) is found in the EDUC row under the "B" column (Unstandardized Coefficients). Its value is reported as -.165. With each additional year of educational attainment, daily television viewing hours decrease by .165 hours on average. Now, we can use the regression line formula to make predictions. For example, those with 16 years of formal education (normative time for a Bachelor's degree) watch

an average of 3.258 hours of television in a typical day [Y = 5.898 - .165(16) = 3.258].

Now, we are going to move to the top Model Summary table. **R square** (r^2) is the proportion of the variation of Y (the dependent variable TVHOURS) that is explained by X (the independent variable EDUC). This model summary statistic is also known as the **coefficient of determination**. By definition, it ranges from 0.0 to 1.0 and is interpreted as a percentage. Thus, we need only multiply it by 100 (or move the decimal two places over to the right) to make it a percentage. The value of R square is reported as .022. Therefore, 2.2% of the variance in TVHOURS is explained by EDUC. This means that 97.8% of the variation in TVHOURS is not explained by EDUC.

The F ratio in the ANOVA table indicates whether the R square of the model is statistically significant and able to be generalized to the population as a whole. We see that the Sig. level is reported as <.001. Thus, we are highly confident that the findings in this model are generalizable. Even though the R square is small (and the unexplained variation large), the model is meaningful. While our level of education does influence the amount of television we watch, there are numerous other factors that influence it as well. Moreover, Americans are not uniform nor consistent in such behaviors.

Pearson's r Correlation Coefficient

The last area of Exhibit 11.9 that was highlighted is the value under the "Standardized Coefficients Beta" column. In a bivariate linear regression, this is the value of **Pearson's r correlation coefficient**. Pearson's r is the appropriate measure of association for scale variables. It is the square root of R square in a bivariate linear regression model. By definition, r ranges from -1.0 to +1.0. Recall Exhibit 11.1 on the extreme possibilities of associations between scale variables. An r of +1.0 is a perfect positive association. An r of -1.0 is a perfect negative association. An r of zero reflects a perfect nonassociation. Of course, these extremes are very rare and we nearly always get a value somewhere in between them. Therefore, we need to use our strength guidelines from Exhibit 8.11 to interpret r as indicating a "weak," "moderate," or "strong" association between the two scale variables.

As you will recall, the "b" in the regression line formula is the slope which indicates the change in the value of Y for each unit change in the value of X. This is what is referred to as the unstandardized slope. These are the raw coefficients reported in the metric of the independent variable. Therefore, there is no set value scale and they can be very small or very large numbers. Pearson's r is also referred to as the **Beta weight** in a bivariate linear regression. Beta weights are standardized slopes that enable us to assess the strength and direction of the association. In Exhibit 11.9, the Beta weight is reported as -.150. According to our strength guidelines (Exhibit 8.11), this Pearson's r indicates that EDUC has a "weak," negative association with TVHOURS.

Quantitative sociologists often begin their analyses of survey data by exploring for statistically significant associations among variables. **Correlation matrices** are a frequently used tool that presents Pearson's r for all possible pairs of the variables entered into the command. In SPSS, the path is Analyze > Correlate > Bivariate. All of the defaults of this dialog are appropriate and we simply move the variables for which we want to compute r into the Variables box. Let's consider the correlation coefficients among TVHOURS, EDUC, AGE, and SIBS. Generate the correlation matrix and confirm that your output matches that in Exhibit 11.10.

Exhibit 11.10. Correlation Matrix of TVHOURS, EDUC, AGE, and SIBS

Correlations

		tvhours hours per day watching tv	educ highest year of school completed	age age of respondent	sibs number of brothers and sisters
tvhours hours per day watching tv	Pearson Correlation	1	-.150**	.192**	.097**
	Sig. (2-tailed)		<.001	<.001	<.001
	N	2683	2648	2475	2641
educ highest year of school completed	Pearson Correlation	-.150**	1	.014	-.202**
	Sig. (2-tailed)	<.001		.391	<.001
	N	2648	3966	3683	3908
age age of respondent	Pearson Correlation	.192**	.014	1	.075**
	Sig. (2-tailed)	<.001	.391		<.001
	N	2475	3683	3699	3650
sibs number of brothers and sisters	Pearson Correlation	.097**	-.202**	.075**	1
	Sig. (2-tailed)	<.001	<.001	<.001	
	N	2641	3908	3650	3968

**. Correlation is significant at the 0.01 level (2-tailed).

The first key to reading a correlation matrix is to understand that the columns and the rows are the same. Therefore, every value of Pearson's r is reported twice. By definition, the values along the diagonal (red line) will always be 1 since any variable is perfectly associated with itself. The diagonal also cuts the matrix into two triangles. The lower and upper halves are mirror images of each other (imagine folding the upper half down to match the lower half). It is common to only report the lower half of a correlation matrix in publications (you may have seen such triangular tables before).

Since our correlation matrix contains four variables, there are six different pairs and values of Pearson's r to consider. For each pair, the value of r, its level of statistical significance, and number of cases (N) for the correlation is presented. As with all of our inferential statistics, we deem Sig. levels ≤ .054 as statistically significant correlations.

If we look at the confluence of the first column (TVHOURS) and the second row (EDUC), we see that Pearson's r correlation coefficient is reported as -.150 with a Sig. level of <.001. Notice that these are the exact same statistics that we found in our linear regression of Exhibit 11.9. EDUC and TVHOURS have a "weak," negative correlation that is highly statistically significant. As respondents increase in educational attainment, they tend to decrease in the amount of television watched.

The next pair of variables is TVHOURS and AGE. This is a "weak," positive correlation that is statistically significant. As age increases, so does television viewing on average. Elderly people who are out of the labor force tend to watch more TV than others. Not only do they usually have more free time, but they may also have financial and physical constraints that prevent them from participating in other activities.[5]

Next up is TVHOURS and SIBS. This is also a "weak," positive correlation that is statistically significant. Those with more brothers and sisters likely spent time watching TV with them when they were growing up.[6] These viewing habits may remain with people throughout their lives. This association could also reflect that larger families may struggle to afford other activities for all of the children.

The next pair is EDUC and AGE. As evident, this association is not statistically significant (educational attainment is similar across various ages). EDUC and SIBS is statistically significant, however. It is a

"moderate," negative correlation. Those with more brothers and sisters tend to have lower educational attainment on average.[7] This could also reflect limited financial means among larger families. Such households may be less likely to be able to fund the college experience for all of the children. Another potential explanation is that some kids from large families get jobs as soon as possible to help support the family. This could also deter college attendance.

The final pair of variables in the correlation matrix of Exhibit 11.10 is AGE and SIBS. This is a "weak," positive correlation that is statistically significant. Older respondents tend to have more siblings on average. This makes sense as sociologists have documented a decline in family size and fertility rates over time.[8] Families are having fewer children than they did decades ago. Thus, younger people are likely to have fewer siblings. As you can see, correlation matrices are quite useful to quickly determine if pairs of scale variables are associated with one another.

Multiple Linear Regression

The final topic of this text is a brief introduction to multivariate modeling. Thus far, our exploration of inferential statistics has been limited to bivariate tests between one X variable and one Y variable. **Multiple linear regression** tests the impact of two or more independent variables upon a dependent variable. The various predictors are typically labeled as X_1, X_2, X_3, etc. Multiple regression is much more powerful and can better represent the complexity of our social world. **Model specification** is the deliberate and careful identification of all of the possible causes of particular attitudes or behaviors. Our literature reviews typically identify multiple independent variables that may influence a dependent variable.

Multiple linear regression is an extension of bivariate regression and employs the same SPSS dialog (Analyze > Regression > Linear). There are several importance differences though. First, the nature of the unstandardized and standardized slopes in regression change when there are multiple predictors. The slopes become **partial slopes** since each is only one of several. The values of the unstandardized slopes (b) in a multiple regression reflect the amount of change in Y for a unit change in a specific X variable while controlling for the other independent (X) variables. Each X variable is only part of the model. Importantly, the Beta weights in a multiple regression are no longer identical to Pearson's r correlation coefficient (a bivariate statistic). The advantage of Beta

weights in a multiple regression model is that they are still standardized to the -1.0 to +1.0 metric and can be directly compared to one another to assess the relative strength of each predictor.

Multiple regression also differs from bivariate regression in that nominal and ordinal variables can be added and used as predictors. We would not use these variable types as X variables in bivariate regression since other techniques (*t* tests and ANOVA) would be more appropriate. To use a nominal variable as a predictor, it must be dichotomous. The nominal variable is typically recoded into a dummy variable with 0 and 1 values. This usually means combining categories and collapsing them into just two. The MINORITY variable that we have employed in this text is a good example.

Since ordinal variables are ranked, it is easier to use them as predictors in multiple regression. The quantitative sociologist must keep in mind that regression only tests for linear associations though. So, an exploratory ANOVA with the ordinal predictor and the scale outcome should be conducted first to see if a linear association exists and if it is appropriate to use that ordinal X variable in a multiple regression model.

The last difference between bivariate and multiple regression is the transformation of R square, the total variation explained. In multiple regression, it is referred to as the **multiple coefficient of determination** and is symbolized with a capital letter (R^2). This statistic reflects the proportion of the total variation in Y that is explained jointly by all of the independent variables.

Let's get to it! Do AGE and EDUC predict internet use (WWWHR)? Use the Analyze > Regression > Linear command and place WWWHR in the Dependent box and AGE and EDUC into the Independent(s) box. Exhibit 11.11 displays the output.

First, we see that AGE and EDUC are both statistically significant predictors of internet use. Age has a "moderate," negative association (Beta weight = -.236). EDUC as a "weak," positive association (Beta weight = .107). Older people use the internet less frequently than younger people (an increase in AGE is associated with a decrease in WWWHR). This seems sensible as the internet has existed for around thirty years and many of the respondents would not have grown up with it. Those with higher educational attainment report using the internet more frequently

than those who are less educated. College students certainly use the internet a lot and many go on to gain employment in fields that require its use. The association could also reflect resource differentials as those with low education may not have the means to afford a computer and/or internet service.

Exhibit 11.11. SPSS Output of Multiple Regression of WWWHR on AGE and EDUC

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.256[a]	.066	.065	17.053

a. Predictors: (Constant), educ highest year of school completed, age age of respondent

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	47117.702	2	23558.851	81.014	<.001[b]
	Residual	672036.064	2311	290.799		
	Total	719153.765	2313			

a. Dependent Variable: wwwhr www hours per week

b. Predictors: (Constant), educ highest year of school completed, age age of respondent

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	17.431	2.236		7.796	<.001
	age age of respondent	-.247	.021	-.236	-11.711	<.001
	educ highest year of school completed	.692	.131	.107	5.295	<.001

a. Dependent Variable: wwwhr www hours per week

The value of R^2 indicates that 6.6% of the variation in internet usage is explained by age and education. The F ratio is highly significant, so we are confident that this model is generalizable to the larger population of American adults.

We've investigated the TVHOURS variable multiple times in this text. Let's look at it one more time using a multivariate model. Do AGE, EDUC, SEXBIRTH1 and MINORITY predict TVHOURS? Generate the model in SPSS and confirm that your output matches that in Exhibit 11.12.

Exhibit 11.12. SPSS Output of Multiple Regression of TVHOURS on AGE, EDUC, SEXBIRTH1, and MINORITY

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.257[a]	.066	.065	3.029

a. Predictors: (Constant), minority people of color (dichotomous race), sexbirth1 r's sex assigned at birth (2021), educ highest year of school completed, age age of respondent

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1587.381	4	396.845	43.252	<.001[b]
	Residual	22396.445	2441	9.175		
	Total	23983.825	2445			

a. Dependent Variable: tvhours hours per day watching tv

b. Predictors: (Constant), minority people of color (dichotomous race), sexbirth1 r's sex assigned at birth (2021), educ highest year of school completed, age age of respondent

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3.720	.438		8.484	<.001
	age age of respondent	.037	.004	.204	10.327	<.001
	educ highest year of school completed	-.169	.022	-.152	-7.745	<.001
	sexbirth1 r's sex assigned at birth (2021)	.122	.124	.019	.986	.324
	minority people of color (dichotomous race)	.524	.155	.067	3.373	<.001

a. Dependent Variable: tvhours hours per day watching tv

AGE, EDUC, and MINORITY are all statistically significant predictors. SEXBIRTH1 is not. There are no sex differences in average hours of television viewing. Age has a "moderate," positive association as older people report watching more TV. EDUC has a "weak," negative association as those with more education watch it less frequently. MINORITY has a "weak," positive association with TVHOURS. Recall that this variable has whites coded as 0 and people of color coded as 1. The positive association means that group 1 is higher on the dependent variable than group 0 (see Exhibit 11.13). In other words, people of color report watching television more frequently than whites. As you may recall, this was documented with a bivariate *t* test back in Chapter 9. The R^2 is statistically significant and indicates that 6.6% of the variation in television viewing hours is explained by the four predictors.

Exhibit 11.13. Dichotomous Predictors and Direction of Association in Multiple Regression Models

	Negative Association	*Positive Association*
Whites	Higher Mean	Lower Mean
People of Color	Lower Mean	Higher Mean

Our last multiple regression model will predict the MEATDAYS variable. This is a new one in the General Social Survey. Respondents were asked to report the number of days in a typical week that they eat red meat (beef or lamb). This scale variable ranges from 0 to 7 days. Do AGE, EDUC, SEXBIRTH1, MINORITY and TVHOURS predict the frequency that one eats red meat (MEATDAYS)? Exhibit 11.14 displays the output.

First, we see that AGE is not a statistically significant predictor. Adults of various ages eat red meat at similar frequencies. Each of the four other variables are statistically significant predictors with "weak" effects upon MEATDAYS. EDUC is the most impactful variable in the model as it has the largest Beta weight value. It is a negative association. Those with more education report eating red meat less frequently. SEXBIRTH1 also has a negative association. Recall that this variable is coded 1 (males) and 2 (females). The negative association indicates that the 2 group is lower on Y than the 1 group (see Exhibit 11.13). In other words, females report eating red meat less frequently than males.

Exhibit 11.14. SPSS Output of Multiple Regression of MEATDAYS on AGE, EDUC, SEXBIRTH1, MINORITY and TVHOURS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.232[a]	.054	.051	1.671

a. Predictors: (Constant), tvhours hours per day watching tv, sexbirth1 r's sex assigned at birth (2021), minority people of color (dichotomous race), educ highest year of school completed, age age of respondent

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	250.155	5	50.031	17.926	<.001[b]
	Residual	4404.199	1578	2.791		
	Total	4654.354	1583			

a. Dependent Variable: meatdays days eating red meat per week

b. Predictors: (Constant), tvhours hours per day watching tv, sexbirth1 r's sex assigned at birth (2021), minority people of color (dichotomous race), educ highest year of school completed, age age of respondent

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.715	.312		15.130	<.001
	age age of respondent	-.003	.003	-.028	-1.109	.268
	educ highest year of school completed	-.097	.015	-.160	-6.397	<.001
	sexbirth1 r's sex assigned at birth (2021)	-.413	.085	-.120	-4.859	<.001
	minority people of color (dichotomous race)	-.449	.106	-.105	-4.219	<.001
	tvhours hours per day watching tv	.047	.015	.082	3.228	.001

a. Dependent Variable: meatdays days eating red meat per week

MINORITY also has a negative association in this model as people of color report eating red meat less frequently than whites. Finally, TVHOURS has a positive association as those who report watching more television eat red meat more frequently. I will leave it to you all to interpret

all of these findings! The R^2 is statistically significant and indicates that 5.4% of the variance in MEATDAYS is explained by the five predictors.

This final chapter concludes our journey into the essentials of quantitative sociology. The ten previous chapters provided the foundation for us to learn about regression techniques to analyze scale variables. Scattergrams plot two scale variables on a graph and enable us to determine if there is a linear relationship between an independent (X) variable and a dependent (Y) variable. The regression line is based upon a simple, but powerful, formula allowing predictions of Y given values of X. The linear regression model also contains the coefficient of determination (r^2), a model summary statistic. Pearson's r correlation coefficient is synonymous with the Beta weight in a bivariate regression. Correlation coefficients are regularly produced in SPSS and the resulting correlation matrices are useful to quickly explore associations (and their direction and strength) among numerous variables. Finally, we learned about the most common type of multivariate modeling, multiple linear regression. The ability to have multiple independent variables predicting a dependent variable enables us to produce models that better reflect our social world and simultaneously test multiple research hypotheses.

Key Terms

Scattergrams, linearity assumption, regression line, residual, case configurations, clustering, regression line formula, Y-intercept, linear regression, R square, coefficient of determination, Pearson's r correlation coefficient, Beta weight, correlation matrices, multiple linear regression, model specification, partial slopes, and multiple coefficient of determination.

Endnotes

1. Data compiled by author from https://worldpopulationreview.com/

2. Hammond, Michael. 2025. "Exponential Expansions in Social Evolution: The Case of Per Capita Income Averages in the U.S.. *Theory and Society*. https://doi.org/10.1007/s11186-024-09589-w

3. Childers, Chandra, Ariane Hegewisch, Tanima Ahmed, and Amy Burnett Cross. 2019. *Geographic Mobility, Gender, and the Future of Work*. Washington, D.C.: Institute for Women's Policy Research

https://iwpr.org/wp-content/uploads/2020/07/C487_Geographic-Mobility-FOW.pdf

4. Christakis, Nicholas A. and James H. Fowler. 2009. *Connected: The Surprising Power of our Social Networks and How they Shape our Lives*. New York, NY: Little, Brown and Company.

5. Depp, Colin A., David A. Schkade, Wesley K. Thompson, and Dilip V. Jeste. 2010. “Age, Affective Experience, and Television Use.” *American Journal of Preventive Medicine* 39 (2): 173-178.

6. Kotler, Jennifer A., John C. Wright, and Aletha C. Huston. 2001. “Television Use in Families with Children.” In J. Bryant & J. A. Bryant (eds.), *Television and the American Family* (2ed., pp. 33-48). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

7. Chihaya, Guilherme and Marta Styrc. 2015. “The Impact of Family Size on Educational Attainment in Cross-Country Comparative Perspective.” Poster presentation at the Annual Meeting of the Population Association of America, San Diego, CA.

8. Fahey, Tony. 2017. “The Sibsize Revolution and Social Disparities in Children’s Family Contexts in the United States, 1940-2012.” *Demography* 54 (3): 813-834.